TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

SINI MOKKILA

CONFORMATIONAL DYNAMICS OF FUNCTIONALLY IMPORTANT LOOPS OF HIV-1 GP120 PROBED BY MOLECULAR DYNAMICS SIMULATIONS

Master of Science Thesis

Examiner: Prof. Ilpo Vattulainen
Examiner and Topic Approved in the
Faculty of Natural Sciences Council
Meeting on 6th May 2015

ABSTRACT

SINI MOKKILA: Conformational Dynamics of Functionally Important Loops of HIV-1 Gp120 Probed by Molecular Dynamics Simulations
Tampere University of Technology
Master of Science Thesis, 66 pages, 13 Appendix pages
May 2015
Master's Degree Programme in Science and Engineering
Major: Technical Physics
Examiner: Prof. Ilpo Vattulainen
Keywords: AIDS, HIV-1, gp120, variable loops, molecular dynamics

Acquired immunodeficiency syndrome (AIDS) is a life-threatening disease caused by human immunodeficiency viruses (HIVs). The availability of antiretroviral drugs has changed AIDS/HIV from a fatal disease to a controllable condition. However, protective vaccination would help to get rid off the plague for good. The main targets for the vaccinal antibodies are surface glycoproteins (gp120) that lie on the viral envelope of HIV-1 in trimeric complexes. Each gp120 unit is characterized by five variable (V) loops. Especially the major variable loops V1, V2, and V3 are functionally important. However, due to their flexibility and variability they are tricky targets and also hard to study experimentally.

Here, atomistic molecular dynamics (MD) simulations are harnessed to improve the current understanding of the gp120 loop dynamics. In this study, all variable loops of gp120 including the most varying loops V1 and V2, and additionally, the native trimeric state of gp120 are considered for the very first time. The sequence of a native HIV-1 isolate, YU-2, was chosen for the study. According to previous MD studies the major variable loop V3 shows significant flexibility in monomeric gp120. This is also observed in this study. Additionally, the flexibility of the V1/V2 domain in monomeric gp120 is demonstrated for the first time. However, gp120 in the trimer context tells a different story. According to the trimer simulation, the major variable loops mostly have lost their flexibility, mobility, and concerted motion among the loops in the trimer. The reduced motion seems to be due to inter-gp120 interactions, where the V2 and V3 loops play key roles in stabilizing the trimer apex.

The results provide significant insight into gp120 variable loop dynamics on an atomic level in the native trimeric state and facilitate understanding the mechanism of the viral entry and its inhibition. However, similar MD simulations need to be repeated in order to confirm the present findings.

TIIVISTELMÄ

SINI MOKKILA: HIV-1 gp120-proteiinin toiminnallisesti merkittävien silmukkarakenteiden dynamiikka molekyylidynamiikan simulaatioiden valossa
Tampereen teknillinen yliopisto
Diplomityö, 66 sivua, 13 liitesivua
Toukokuu 2015
Teknis-luonnontieteellinen koulutusohjelma
Pääaine: Teknillinen fysiikka
Tarkastajat: Prof. Ilpo Vattulainen
Avainsanat: AIDS, HIV-1, gp120, muuntautumiskykyiset silmukat, molekyylidynamiikka

Immuunikato (AIDS) on immuunikatovirusten (HIV) aiheuttama hengenvaarallinen sairaus. Nykyään AIDS/HIV-potilaiden tilaa voidaan kontrolloida antiretroviraalisilla lääkkeillä, mutta virukselta suojaavan rokotteen avulla tästä ikävästä taudista päästäisiin lopullisesti eroon. Kehitteillä olevien rokotteiden pääkohde on virusvaipan pinnalla lymyilevät gp120-glykoproteiinit, jotka muodostavat keskenään trimeerisiä komplekseja. Kunkin gp120-proteiinin rakenteeseen kuuluu viisi muuntautumiskykyistä silmukkarakennetta. Näistä etenkin kolme suurinta silmukkaa (V1, V2 ja V3) ovat tärkeitä proteiinin toiminnan kannalta, mutta niiden tutkiminen kokeellisesti on hankalaa, koska ne ovat joustavia ja niiden proteiinisekvenssi vaihtelee.

Tässä tutkimuksessa perehdytään gp120-proteiinin silmukoiden dynamiikkaan atomitason molekyylidynaamisten (MD) simulaatioiden avulla. Kaikki gp120:n viisi silmukkaa on tässä työssä ensimmäistä kertaa MD-simulaatioiden historiassa onnistuttu lisäämään proteiinin rakenteeseen. Lisäksi gp120:n trimeerinen rakenne on toteutettu simuloitavaksi ensimmäistä kertaa. Proteiinin sekvenssi otettiin HIV-1 tyypin viruksen isolaatista, YU-2:sta. Edellisten MD-simulaatioiden perusteella vaikuttaisi siltä, että V3-silmukka on poikkeuksellisen joustava rakenne monomeerisen gp120:n pinnalla. Tässä tutkimuksessa saatiin vastaavia tuloksia. Lisäksi tässä tutkimuksessa osoitetaan ensimmäistä kertaa, että myös silmukat V1 ja V2 ovat poikkeuksellisen joustavia gp120-monomeerissä. Osoittautuu kuitenkin, että trimeerisessä kompleksissa gp120:n silmukat menettävät merkittävästi joustavuuttaan ja liikkuvuuttaan. Myös V1, V2 ja V3 silmukoiden välillä havaittu kollektiivinen liike vähenee trimeerissä verrattuna monomeeriin. Kyseiset dynaamiset erot aiheutunevat gp120-yksiköiden välisistä vuorovaikutuksista, missä etenkin V2- ja V3-silmukat ovat aktiivisessa roolissa, tasapainottamassa trimeerin ulointa osaa.

Tässä tutkimuksessa saadut tulokset vaikuttavat keskeisesti ymmärrykseemme siitä, miten gp120:n silmukkarakenteet käyttäytyvät atomitasolla niiden luonnollisessa tilassa trimeerikompleksissa, ja auttavat ymmärtämään viruksen tunkeutumista soluun ja kyseisen tapahtuman torjumista. Vastaavia MD-simulaatioita tarvitaan kuitenkin lisää varmistamaan saatuja johtopäätöksiä.

PREFACE

This Master of Science Thesis was carried out in the Biological Physics and Soft Matter (BIO) Group at Tampere University of Technology. The computing facilities were provided by the Partnership for Advanced Computing in Europe (PRACE) association and the Finnish IT Center for Science (CSC).

I have had great pleasure to work in the BIO Group already since 2012. Firstly, I want to thank my employer and Thesis examiner Ilpo Vattulainen for keeping me busy with various intriguing projects and for always encouraging me to put my best foot forward. I also want to thank Moutusi Manna for supervising me with expertise, warm support and vigilance. Finally, I want to thank my other coworkers for sincerely helping me in my everyday work, lunching and having hilarious dinners.

In fact, the whole time in my university has been a blast. Voluntary work in student organizations has provided me with valuable skills in organizing and managing, but most importantly, in having fun. This is all thanks to my "teekkari" friends and acquaintances that were there as well, and I cannot wait to meet them again in my future career.

Tampere, 16th April 2015

Sini Mokkila

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AIDS	Acquired immunodeficiency syndrome
Cα	α-carbon
C1	Conserved domain 1
C2	Conserved domain 2
C3	Conserved domain 3
C4	Conserved domain 4
C5	Conserved domain 5
CCR5	C-C chemokine receptor type 5
CD4	Cluster of differentiation 4
CDR	Complementary determining region
COM	Center of mass
CXCR4	C-X-C chemokine receptor type 4
Env	Envelope glycoprotein
FP	Fusion peptide
gp120	Glycoprotein 120
gp160	Glycoprotein 160
gp41	Glycoprotein 41
H-bond	Hydrogen bond
HIV	Human immunodeficiency virus
HIV-1	Human immunodeficiency virus type 1
HIV-2	Human immunodeficiency virus type 2
HR	Heptad repeat
HR1	Heptad repeat 1
HR2	Heptad repeat 2
HXB2	HIV-1 isolate
LJ	Lennard-Jones
MD	Molecular dynamics
MPER	Membrane proximal external region
NMR	Nuclear magnetic resonance
NpT	Normal pressure and temperature
NVE	Normal volume and energy
NVT	Normal volume and temperature
PC	Principal component
PC1	Principal component 1
PC2	Principal component 2
PC3	Principal component 3

PCA	Principal component analysis
PDB	Protein Data Bank
PBC	Periodic boundary conditions
PME	Particle mesh Ewald
R_G	Radius of gyration
RMSD	Root-mean-square deviation
RMSF	Root-mean-square fluctuation
SIV	Simian immunodeficiency virus
SPEM	Smooth particle mesh Ewald
TM	Trans membrane domain
v-rescale	Velocity rescaling
V1	Variable domain 1
V2	Variable domain 2
V3	Variable domain 3
V4	Variable domain 4
V5	Variable domain 5
YU-2	HIV-1 isolate

1. INTRODUCTION

Acquired immunodeficiency syndrome (AIDS) is a life-threatening disease that has been a worldwide health problem and a plague for human society [1]. It is caused by a retrovirus termed human immunodeficiency virus (HIV) [1]. Two types of HIV have been characterized, the type 1 (HIV-1) and 2 (HIV-2), of which the aforementioned is the cause of the majority of HIV infections globally [2]. The viruses can further be divided into groups, subtypes, and finally isolates that have diverged in infected individuals over time. Typically, certain isolates appear in greater extent at particular geographical locations.

The HIV-1 displays trimeric envelope glycoproteins termed *Envs* on its surface [3]. Their structural information is of critical importance for rational vaccine design and for understanding the detailed mechanism of viral entry [3]. Env is a noncovalently bound complex of three heterodimers consisting of transmembrane glycoproteins (gp41) and surface glycoproteins (gp120) [3]. The gp120 mediates attachment of the virus to the target cell, whereas gp41 attaches gp120 to the viral membrane and is required for the fusion of the viral and target cell membranes [4]. So far, numerous crystal structures of these Env subunits as well as of the complete trimer have been reported [3]. However, some details from the structure are still missing, especially from the less studied native trimeric conformation. Additionally, the numerous structural studies which are the starting point and the ultimate basis of every protein study, lack an essential character often needed to describe protein function, the dynamic aspect.

The gp120 subunits of Env are the main target for neutralizing antibodies and hence of great interest in HIV-1 study. Gp120 has a relatively rigid, conserved, and well described core sructure within various HIV-1 isolates. Additionally, it has five flexible and, among different isolates, highly varying loop structures that are well exposed on the protein surface [5]. The crystallization of these *variable (V) domains* or *loops* appeared to be a challenging task, and hence they have often been unresolved in or excluded from the structural determination. However, due to their high flexibility this kind of regions generally have fundamental roles in determining the protein dynamics, such as large concerted motions, conformational transitions as well as the

recognition and binding of receptors [6]. There are three major variable loops (*V1*, *V2*, and *V3*) in one gp120 which are known to play important functional roles and significantly affect the gp120 dynamics. Additionally, there are two shorter variable loops (*V4* and *V5*). However, only the dynamics of the V3, V4, and V5 loops have been covered in previous molecular dynamics simulation studies [6, 7]. What is more, in these studies only the dynamics of gp120 as a *monomer* have been studied, even though the native state of gp120 is a *trimer*.

This study aims at answering these shortcomings by using atomistic molecular dynamics (MD) simulations. MD is a novel computational method that can be used to complement experiment-based structural studies, such as nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy, and X-ray crystallography. MD is based on classical Newtonian mechanics where the motion of atoms is described by their nuclei. When the crystal structure of a protein is known, it is possible to run MD simulations that imitate the real-life behavior of the protein. The simulation may be then visualized and analyzed with various computational tools. In this way, events in the atomic scale become understandable to human eyes and real quantities from the protein can be measured. MD often enables experiments that are not even possible or are too expensive to carry out in practice. Additionally, instead of a static view that is gained in many experimental methods, MD shows the dynamics, the time evolution of the structure. Hence, MD has power to support, explain, and predict experimental observations.

This study concentrates on examining the conformational dynamics of the HIV-1 envelope glycoprotein gp120 both in monomer and trimer conformations. All variable loops (V1 to V5) are included in the gp120 structure. Thus, the improvements to previous studies are the inclusion of the variable loops V1 and V2, and the native trimeric state. Special emphasis is laid on the major variable loops, V1, V2 and V3. The structure of this Thesis is the following. Chapter 2 introduces the underlying biology. Chapter 3 covers the methodology of molecular dynamics simulations. Chapter 4 introduces the simulation systems and the protocol followed here. Chapter 5 covers the results and discussion. Finally, Chapter 6 concludes the most important findings.

2. BIOLOGICAL BACKGROUND

In this Chapter an overview of AIDS and its cause HIV is given. First, historical and general knowledge of the disease are provided. Then, the structure and functions of the most relevant components on the virus are presented. Finally, all parts are put together in the context of viral entry, that is, how the virus recognizes and penetrates its target cell in order to infect it.

2.1 Past Three Decades of AIDS

AIDS is a life-threatening disease that is induced by HIVs and simian immunodeficiency viruses (SIVs) in their respective human and simian hosts [1, 8, 9, 10]. HIV spreads horizontally, that is, from one individual to another by sexual routes and blood contact [1]. The lethality of the disease comes from the pronounced depression of cellular immunity [1]. The virus targets cells that play central roles in the immune system such as T lymphocytes, monocytes, dendritic cells, and brain microglia [11]. These target cells are characterized by CD4 surface glycoproteins [11].

There are two types of HIV with similar virological properties, the type 1 (HIV-1) and 2 (HIV-2) [2]. Early investigations following the identification of HIV have indicated the type 1 to be more infectious and capable of inducing AIDS than the type 2 [2]. Thus, it has awoken more interest on research fields as well as in this study. The strains of HIV-1 can further be classified into four groups, the *major* group M, the *outlier* group O and two other groups, N and P [12]. Most HIV-1 infections are globally caused by group M viruses [13]. Within group M there are plenty of subtypes or clades. They are designated by the letters A-D, F-H, J and K [12], see Figure 2.1. In Europe the subtype B is the most common, and also comprises 10 % of the HIV-1 infections worldwide [13]. Finally, the subtypes may further be divided into isolates.

An estimated 60 million people have been infected with HIV-1 since the beginning of the AIDS pandemic and approximately half of them have died [14]. According to estimates of Joint United Nations Programme on HIV/AIDS more than 30 million people are currently infected with HIV with the most of them living in sub-Saharan

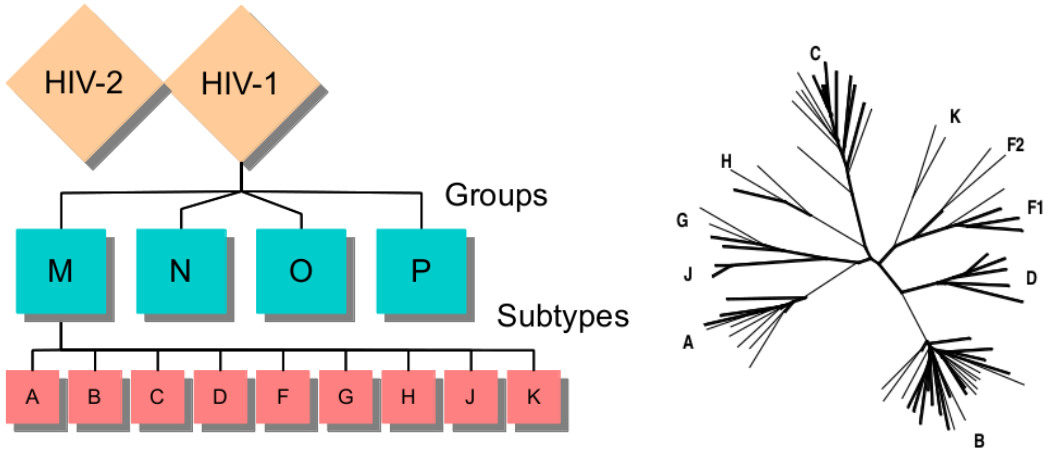


Figure 2.1 *The groups and subtypes of HIV-1 and the phylogenetic relationships between representative strains of group M subtypes A-D, F-H, and J-K from env (gene) nucleotide sequence comparison. The picture on the right is from Reference [12].*

Africa [14]. Despite the educational campaigns about HIV transmission and the fact that highly effective antiretroviral drugs became available in the mid-1990s still in some regions, such as Kazakhstan and Sri Lanka, the incidence of HIV infections is still increasing [14]. The availability of highly efficient antiretroviral drugs has significantly improved the prospects for patients with HIV infection [14]. Especially in the industrialized world antiretroviral drugs are widely available, and thus HIV/AIDS has changed from a rapidly fatal disease to a chronic, controllable condition [14]. However, the development of protective vaccination has made little practical progress [14], and to this day there is no known complete cure for AIDS [14]. One reason for this is that HIV is a rapidly mutating virus exhibiting a range of genetically diverse strains [15]. Additionally, the main targets for antibodies on the viral envelope of HIV-1 are masked by flexible protein loops and glycans [4]. The dynamics of these structures are poorly understood.

2.2 Characteristic Viral Envelope of HIV-1

HIV-1 is an envelope virus with dimensions of about 100 nm to 150 nm [17]. A schematic picture of an HIV-1 virion is shown in Figure 2.2. Envelope viruses are covered with viral envelopes derived from portions of the host cell membranes. The envelopes are essential for the entry into host cells [18]. Specific glycoproteins on their surface serve to identify and bind to receptors on the host [18]. The glycoproteins are composed of two kinds of glycoprotein (gp) subunits, gp41 and gp120 [19]. The subunits form noncovalently linked heterodimers, gp41-gp120, that each further bind noncovalently to two other heterodimers [20]. Hence, the mature envelope

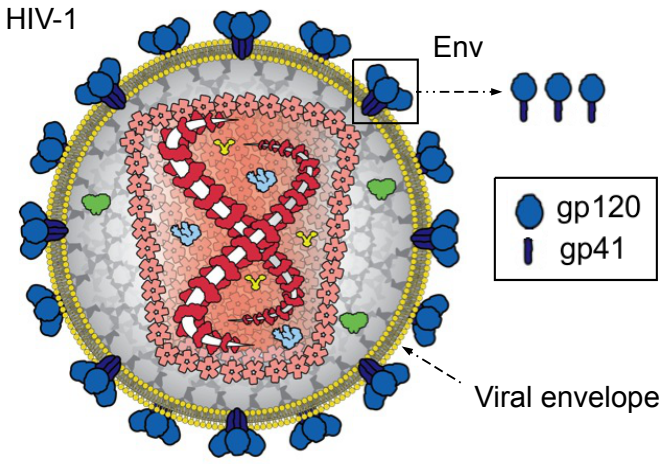


Figure 2.2 *Schematic picture of the HIV-1 virion. The viral envelope is characterized by trimeric structures termed Envs that are composed of gp120 and gp41 subunits. The picture is modified from Reference [16].*

glycoprotein termed *Env* is a heterotrimer. For Envs, see again Figure 2.2. Gp120 and gp41 originate from a same precursor gp160. In the Golgi, gp160 is extensively glycosylated and proteolytically cleaved into gp120 and gp41 [20]. For the sequence and cleavage, see Figure 2.3. Envelope glycoproteins are of great interest as inhibiting their functions would directly disturb the viral fusion, and hence prevent the virus from spreading. Next, the Env subunits gp41 and gp120 are presented more closely. In this study gp120 is of interest and more extensively discussed. The topic has also been discussed in various reviews [4, 11, 19, 21].

2.2.1 Transmembrane Protein Gp41

The transmembrane protein gp41 is a subunit of the Env trimer. Its structure has been extensively studied [11, 22, 23, 24, 25]. Gp41 consists of about 300 residues that form an extracellular domain and a transmembrane domain. The extracellular domain is composed of a fusion peptide, two helical heptad repeats (HR), a loop region, and a membrane proximal external region. For the sequence of gp41, see Figure 2.3. Gp41 subunits attach gp120 subunits to the viral membrane and play key roles in the fusion of viral and host-cell membranes. Gp41 is known to undergo major conformational changes during the fusion discussed in more detail in Section 2.3. Two conformational states before and after the fusion are shown in Figure 2.4.

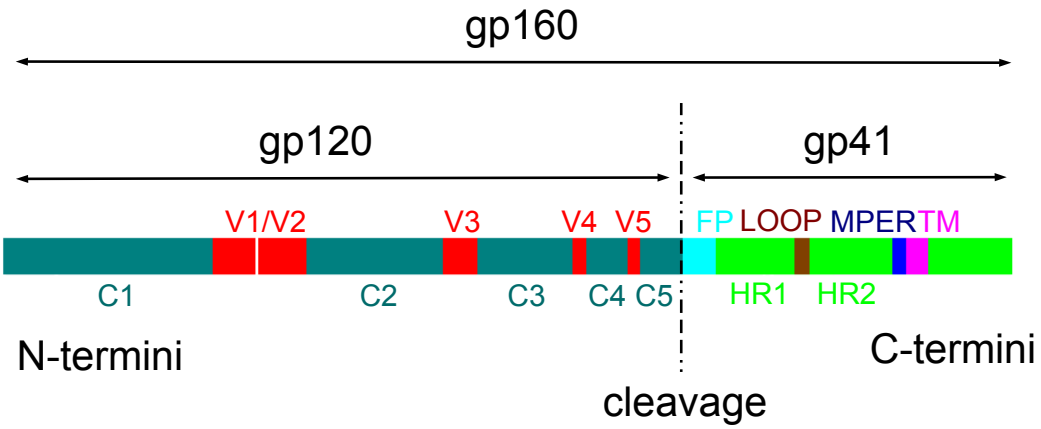
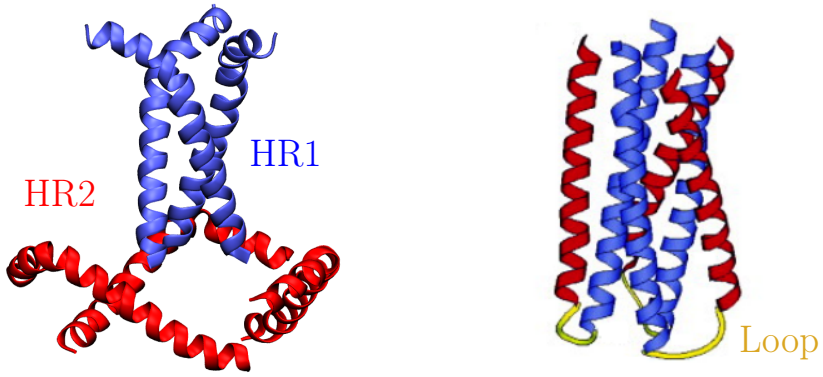


Figure 2.3 *Sequence and cleavage of the precursor gp160 into the gp120 and gp41 subunits. The sequence of gp120 can be divided into five conserved (C) and five variable (V) domains. The sequence of gp41 comprise the fusion peptide (FP), two heptad repeats (HR1, HR2), the loop region, the membrane proximal external region (MPER), and the transmembrane (TM) domain.*

2.2.2 Surface Protein Gp120

The surface protein gp120 is the other subunit of the Env trimer. In contrast to gp41, it lies exterior to the membrane, and hence forms the apex of the viral envelope. Based on its sequence, gp120 is divided into five *conserved* (C) and five *variable* (V) regions [26]. For the sequence, see Figure 2.3. The conserved domains fold into a gp120 core [19], whereas the variable domains are well exposed loops on the surface of gp120 [27, 28, 29]. The first crystal structure of the gp120 core came out in the late 1990s [19], whereas the characterization of the variable loops appeared to be challenging and took much longer. After the first crystal structure, various crystal structures of gp120 cores have been published [30, 31, 32]. Crystal structures of proteins are often derived in complex with other molecules or parts of them, such as antibodies and receptors that they bind to. The binding usually stiffens the protein structure, which in turn enhances the resolution. Accordingly, gp120 is often crystallized in a complex with its primary receptor CD4 in the *CD4-liganded state*, and with specific antibodies. For such bound state, see Figure 2.7.

The overall conformation of the gp120 core in all resolved crystal structures is practically the same. Typically the *inner* and *outer* domains are clearly separated, the nomenclature indicating the expected orientation of the domains within an Env. The inner domain faces the heart of the trimer, whereas the outer domain is mostly exposed on the surface. Additionally, there is a β-sheet domain termed a *bridging sheet* that forms a link between the inner and outer domains. For the domains, see



(a) Pre-fusion gp41 (b) Post-fusion gp41

Figure 2.4 *A schematic picture of two conformational states of the trimeric gp41. A pre-fusion and a post-fusion state are shown. The picture (a) represents the crystal structure in Reference [33]. The picture (b) is modified from Reference [34].*

.

Figure 2.5. However, since the gp120 core remains unchanged even if it is bound to different antibodies and receptors, possible conformational changes and functional effects caused by them are to be seen somewhere else than in the core conformation. An example of such a functional effect is the inhibiting effect of antibodies. Instead of the core, in gp120 these kinds of effects may, however, lie in minor differences in the binding site, in quaternary conformational changes, and/or in the variable domains that are not easily accessible to crystallization [17].

The main function of gp120 is to direct target cell recognition and viral tropism [11]. Gp120 binds both to CD4 glycoprotein receptors on the target cells and to chemokine coreceptors. The binding site for CD4 on gp120 lies in the interface between the inner domain, bridging sheet, and outer domain [19, 36]. The coreceptor-binding site on gp120 lies in the vicinity of the V3 loop [37, 38]. Gp120 is also the main target for neutralizing antibodies and thus in developing vaccines against HIV [4]. However, gp120 is a challenging target due to its heavy glycosylation and variable loops [4]. Glycosylation is a posttranslational modification, where glycans are covalently attached to a protein [39]. Glycans are compounds of glycosidically linked monosaccharides. Both variable loops and glycans are flexible and varying and hard to crystallize [4]. Additionally, they extensively mask potential binding epitopes on gp120 [4]. The glycosylation of gp120 is shortly discussed next and the variable loops in more detail in the next section.

Glycans are known to be crucial for many protein features and functions in cells [40]. They comprise about half of the total mass of gp120 [41, 42]. The inner domain is largely devoid of glycans, whereas the outer domain is mostly covered by them

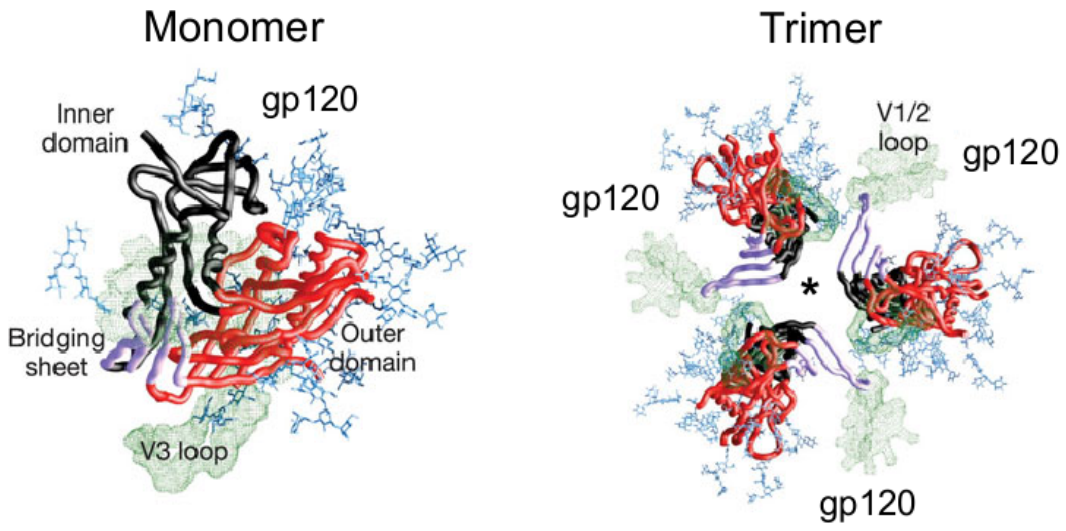


Figure 2.5 *On the left, a schematic gp120 monomer is shown. The viral envelope would be located below it. The core of gp120 is composed of the outer domain, the inner domain, and the bridging sheet. On the right, gp120 is shown in its trimeric conformation. The approximate location of the major variable loops, V1/V2 and V3, and some illustrative glycans masking the gp120 outer domain are also shown. The Env core formed by gp41 heptad repeats is not shown but the location is pointed out with a star (*). The picture is modified from Reference [35].*

[19, 36]. In gp120 glycans are thought to be important for the stabilization and correct folding of the protein [43, 44], and they have been shown to increase the binding affinity of gp120 for CD4 [45]. Additionally, the extent of glycosylation has been linked to specificity of gp120 [38, 46]. What is more, glycosylation hinders the antibody recognition [31]. In fact, so called *glycan shielding* [47] is one of the mechanisms through which viruses have evolved to escape immune system. This is due to the fact that carbohydrate-protein interactions are typically weak [48]. However, nowadays a variety of antibodies are also known that directly bind to the HIV glycan coat [31, 49].

2.2.3 Variable Domains

Variable domains or loops of gp120 are named for their varying sequences. They contain extensive amino acid substitutions, insertions, and deletions among various viral isolates [41]. Despite the variations, the loops are important determinants or indicators for coreceptor and antibody specificity, virus pathogenesis, and disease progression [47, 50, 51, 52]. The first four variable regions, V1 to V4, form surface exposed loops that contain disulphide bonds at their base [41]. The loops V1 and V2 are adjacent and often regarded as a joint region (V1/V2) that comprises the

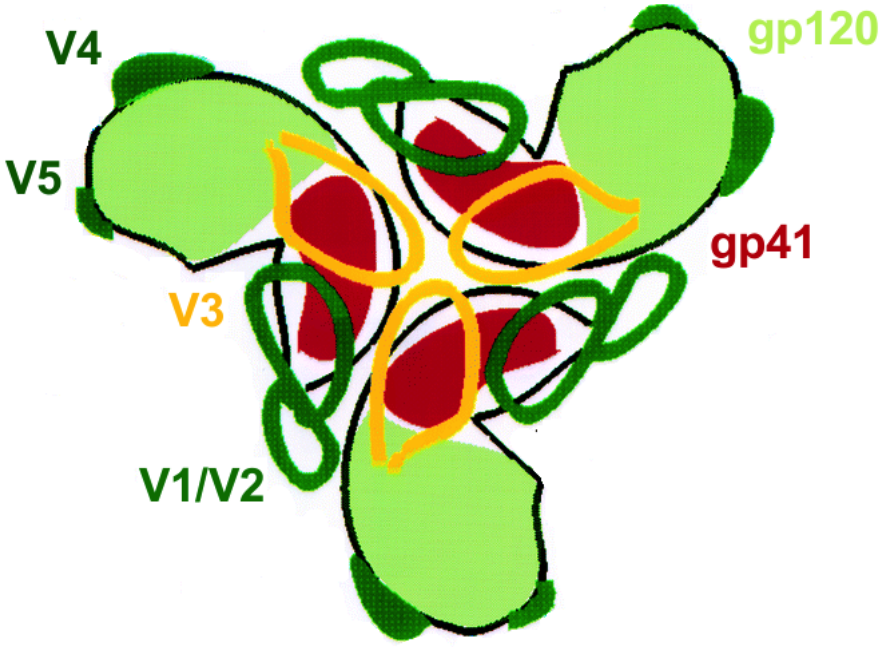


Figure 2.6 *Schematic picture of variable domain locations. The major variable domains, V1/V2 and V3, locate on the apex of Env. The picture is modified from Reference [36].*

most extent variable domain of gp120. Additionally, V3 is an extent loop, whereas V4 and V5 are relatively small. The V1/V2 and V3 loops are often regarded as the most important variable regions in the context of gp120 functions. The location of the variable loops are shown in Figure 2.6.

The V3 domain first characterized in atomic-level in 2005 is a single loop that typically consists of 31 to 39 residues [30]. The loop emanates from the outer domain of gp120 and is almost 50 Å long from the disulphide bridge at its base, see Figure 2.7. Structurally V3 can be divided into three regions: a conserved base, a flexible stem, and a β-hairpin tip [30]. The V3 loop plays the key role in co-receptor binding and specificity [51]. Additionally, V3 is an important factor in determining the overall sensitivity of the virus to neutralization [28]. That is, the loop masks conserved domains on gp120 that would otherwise be ideal binding sites for neutralizing antibodies [28].

The V1/V2 domain, comprised of about 50 to 90 residues, is the most varying domain of gp120 and highly glycosylated [53]. It lies on the apex of gp120, and is of a key role determining the overall sensitivity of gp120 to neutralization [28]. The V1/V2 domain resisted atomic-level characterization for long despite extensive effort. Not until 2011, the first accurate crystal structure of gp120 with the V1/V2 domain came

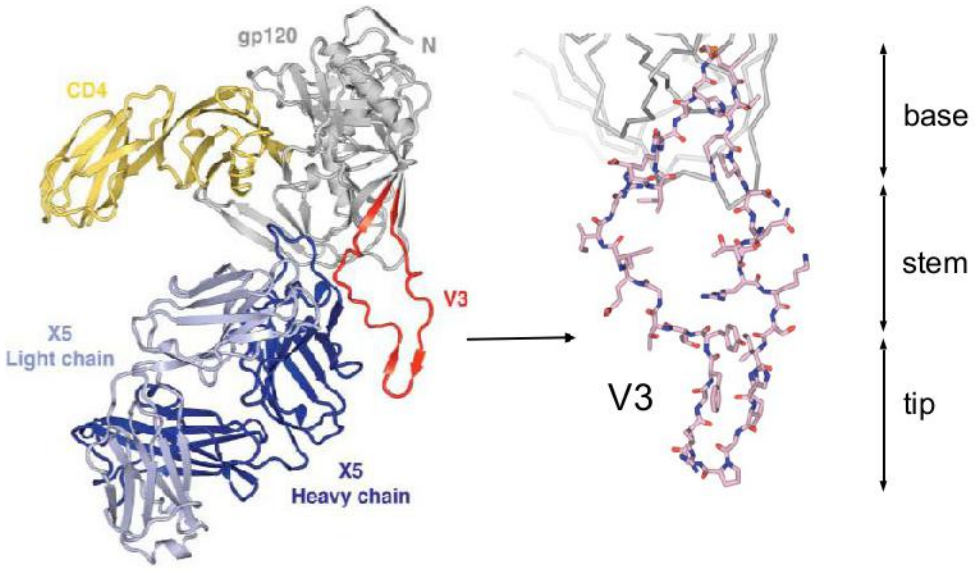


Figure 2.7 *Disposition and structure of the variable loop V3. On the left, the crystal structure of gp120 core (gray) with a V3 loop (red) is shown. A part of the CD4-binding epitope (yellow) and the light and heavy chains of the X5 antibody (two shades of blue) that bind to gp120 are also shown. On the right, the structure of the V3 loop is presented. The pictures are from Reference [30].*

out [53]. This structure is shown in Figure 2.8. The V1/V2 domain forms a four-stranded β-sheet domain. The strands are designated A, B, C and D. The strands A and B form the V1 loop, and the strands C and D the V2 loop. The strands mostly bind to each other by inter-strand disulphide bonds and hydrogen bonds. Additionally, according to the authors [53] the loops V1 and V2 do not just form a continuous sequence but share important structural elements, such as a hydrophobic core and disulphide bonds crossing the strands. Thus, the domain should in their opinion be structurally considered as a single topological entity.

However, the V1/V2 apparently undergo major conformational changes during the CD4 binding [54, 55, 56]. Thus, differences in the conformation of the V1/V2 domain in regard to the above-described may well arise. In a *trimer* ensemble used to build the model in this study [33] the four-stranded β-sheet structure is partly lost, for instance. A more recent *monomer* structure [57], in turn, again suggests it to be present. The latter structure has more resolved residues, however, it only describes a monomer and not the native trimer conformation. Nonetheless, the newest structure came out too late in regard to this Thesis project, and hence is not part of it but to be considered in the future.

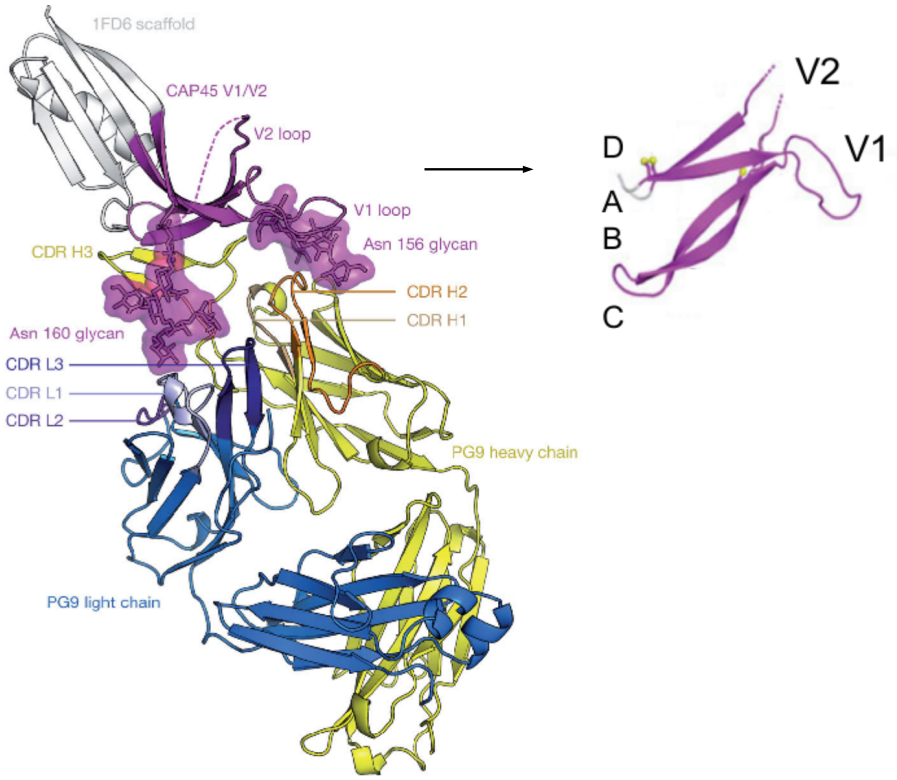


Figure 2.8 *The structure of the V1/V2 domain of gp120 in complex with the PG9 antibody. On the left, the overall structure of V1/V2 and its binding to GP9 are shown. The V1/V2 domain (violet) emanates from gp120 (white). Two glycans at the residue positions 156 and 160 are also shown. The viral membrane is positioned toward the top of the page. The light and heavy chains of PG9 (yellow and blue) and their complementary determining regions (CDRs) are pointed out. CAP45 refers to an HIV-1 strain. On the right, the V1/V2 domain structure is shown alone. The β-strands are labeled A, B, C, and D. The pictures are modified from Reference [53].*

2.3 Viral Entry of HIV-1

In the previous section Env trimers mediating the recognition of and fusion into the target cell were presented in detail. Especially, the structure and functions of the subunits gp120 and gp41 were discussed. In this section these parts are put together in order to get an overview of viral entry. Interestingly, the viral fusion of HIV-1 strikingly reminds that of other enveloped viruses, such as influenza and Ebola [58]. In these viruses, a precursor is derived and cleaved into subunits that form trimeric envelope glycoproteins on the cell surface, similarly to the derivation and cleavage of gp160 into gp120 and gp41 and the formation of the Envs. Thus, understanding the mechanism of HIV-1 virus entry possibly serves to establish a general model for viral membrane fusion of several envelope viruses.

HIV-1 delivers its genetic material into the cell by direct fusion of the viral membra-

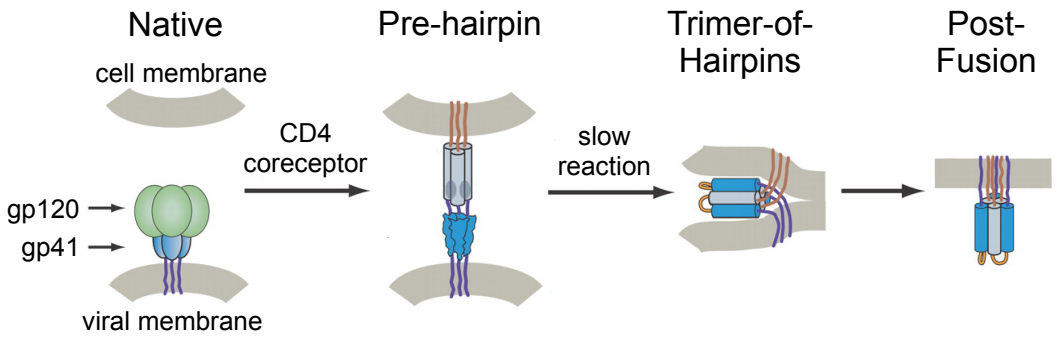


Figure 2.9 *A current model of the viral entry. First, Env is in a native state. The complex triggering mechanisms that activate Env involve the target cell CD4 and coreceptors. The activation results in a prehairpin-intermediate state. A slow reaction results in the trimer-of-hairpins conformation. The final state is the post-fusion state. The picture is modified from Reference [61].*

ne with the cell membrane of the host [59]. A current model of the viral fusion is demonstrated in Figure 2.9. For the sake of clarity, only one Env is shown even though several of them are needed for an efficient fusion to take place [21]. In the *native* state the viral membrane is characterized by the Env glycoproteins, where the gp120 spike masks the gp41 subunits that root Envs to the viral membrane and contain the machinery for the fusion with the host cell membrane [21, 60]. The triggering mechanisms that activate Env from this native state are rather complex involving target cell CD4 receptors, coreceptors, and possibly other cell surface components [21, 60].

First, gp120 binds to CD4 [62, 63] and its coreceptor on the host membrane [64, 65, 66]. A major function of the CD4 binding is to initiate conformational changes in gp120 that contribute to the formation or exposure of the binding site for the coreceptor [67, 68]. Additionally, the CD4 binding promotes conformational changes in gp41 [21]. The function of the coreceptor binding, in turn, has been thought to induce further conformational changes in the envelope glycoprotein complex [21]. Most HIV-1 primate isolates use CCR5 coreceptor and later in the course of infection CXCR4 coreceptor in addition to CCR5 [66, 69]. The preference for one over the other has been defined to arise from the amino acid sequence of the variable loop V3 [51].

However, in the final conformation the gp41 fusion peptide becomes exposed and is inserted into the target membrane which results in the formation of a *pre-hairpin* structure [60]. After this, a further slow reaction results in membrane apposition,

trimer-of-hairpins formation [60]. The interactions crucial for the fusion are unknown but may involve aggregation of gp41 trimers to form fusion pores [21, 60]. Finally, the structure adapts the *post-fusion* state. In this final conformation the fusion peptide and the transmembrane segment of gp41 lie within the same membrane forming a six-helix bundle [70]. In the meantime, the viral genome has been released to the target cell where it may attach to the host genome. The infected host starts to produce new viral particles that are further gathered and budded from the cell membrane as new virions.

3. MOLECULAR DYNAMICS SIMULATIONS

This study aims at understanding possible functions arising from the conformational dynamics of the HIV-1 envelope glycoprotein gp120. For this purpose a novel computational method, molecular dynamics (MD), was applied. MD is a computational simulation technique that represents the interface between theory and experiment [71]. Simulations in general serve to imitate real-world systems over time. Today, MD is recognized as a method that offers important tools for understanding the physical basis of the structure and function of biological ensembles [72, 73]. The method is often used to interpret experiments and to complement them, but also to try out something new that is perhaps too expensive or impossible to carry out with current experimental methods [72]. It has especially shed light on cellular membrane research, that is, on membrane proteins carrying out vital functions together with their modulators, lipids [74]. What is more, proteins are no more regarded as rigid structures, instead their dynamic nature, internal motions, and resulting conformational changes are seen more and more relevant [72]. MD is a great research method for studies of such soft dynamic structures.

The simulations for this study were carried out by using the GROMACS software package [75]. In this Chapter an overview of the methodology is given. The Chapter is based on the manual of GROMACS package [75] and References [76, 71, 77]. First, design and preparation of the systems, and the underlying physics are covered. Next, setting up experimental conditions, temperature and pressure, are presented. Then, how an originally finite simulation system is handled as an infinite lattice, is covered. Then, running MD simulations is presented, and the analysis part is shortly discussed. Lastly, the limitations of MD simulations are discussed.

3.1 Structure Defines Atomic Positions

An overview of running molecular dynamics simulations is shown in Figure 3.1. In this Section, preparation the system is discussed. Firstly, one needs to consider the research problem. Based on it, biologically relevant systems are designed. Things to consider are the number of different macromolecules (proteins, lipids, carbohy-

drates, nucleic acids) and the solution where they naturally are dissolved in (water and its ion concentration). One aims to get a possibly natural imitation, however, simplifications are required due to limited computational resources and expenses, but also, to keep the system understandable or manageable. In order to study a membrane bound receptor, for instance, one does not need to take the whole cell with its membranes and thousands of receptors but rather a tiny patch of the main membrane with one or a few receptors might well be sufficient. The studied system is originally finite but actually becomes infinite by means of periodic boundaries discussed in Section 3.4.

Next, the desired 3-dimensional structures need to be found in databases such as the RCSB Protein Data Bank (http://www.rcsb.org/) for proteins and nucleic acids. The Protein Data Bank (PDB) is a textual file format describing the atomic positions and connectivity. The structural data is gained in experimental crystallographic or nuclear magnetic resonance (NMR) studies, for example. However, the structures might still lack some crucial information. Frequently, protein sequences are incomplete, and flexible and varying parts of the proteins, such as loops and glycans, are missing from the structure. Additionally, experimental procedures applying antibodies and unnatural solutions might have caused changes in the structure. Many structures are provided in their bound conformations, for example, which make them more rigid, and hence easier to crystallize. In these structures there are receptors and antibodies bound to proteins or parts of them, see Figures 2.7 and 2.8. In order to prepare the desired systems, one often needs to remove what is not relevant, and on the other hand, add what is missing. There are special programs to assist the preparation. However, choosing the most relevant structures and the preparation often requires good understanding of the underlying biology and many trials and errors when trying to fit all parts together.

3.2 Force Field Describes Underlying Physics

When the system has been adequately chosen and prepared, it is time to consider the physical laws and parameters governing it. For this, a suitable *force field* is chosen and described in the *topology*. Force field includes the mathematical formulas and parameters that are used to describe the potential energy experienced by atoms. Potential energy is the energy that an atom has only due to its position. It is based on two kinds of interactions, *bonded* and *non-bonded*. The total potential energy function is the sum of the two,

$$V_{total} = V_{bonded} + V_{non-bonded}. \tag{3.1}$$

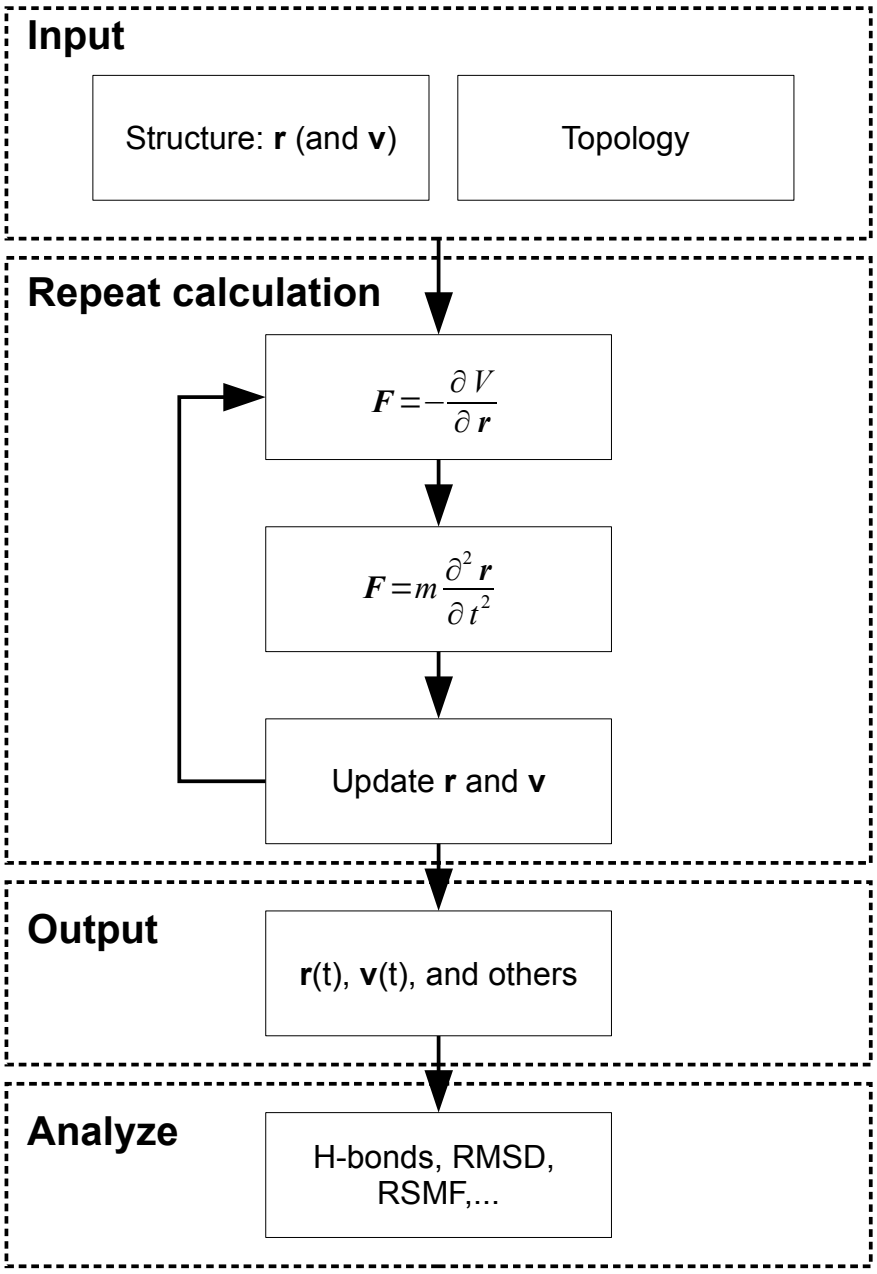


Figure 3.1 *General algorithm for molecular dynamics simulation. First, the structure is needed to describe atomic positions and optionally the initial velocities of the system. The topology describes the underlying physics, such as masses, charges, and bonded and non-bonded interactions. Second, calculations are performed. The same calculation pattern is repeated again and again. As a result, new coordinates and velocities are gained at each step. Third, the output describes the coordinates, velocities, and other desired quantities over time. Finally, analyses are carried out.*

However, not only the potential energy but also the forces are of interest. Mathematically forces are easily derived from the potential energy functions. Additionally, special potentials from fixed lists are used to impose restraints on the motion of the system [75]. Different kinds of restraints, such as position and orientation restraints, are used to avoid destructive deviations or to include knowledge from experimental data [75].

Force fields have been developed by using high level quantum calculations or fitting to experimental data [75]. There are numerous force fields for different purposes. They do not usually belong to the simulation packages themselves, but compatibility between those two is required [75]. Popular force fields specifically designed for atomistic simulations are AMBER, CHARMM and OPLS/AA [78]. In this study OPLS/AA was used. Typically, the potential energy functions between different force fields are similar to each other. Instead, the parametrizations may vary significantly. In the next sections, some general formulation of the potential energy functions in force fields are described.

3.2.1 Bonded Interactions

Bonded interactions describe the interplay between atoms that are covalently linked to each other. They are based on fixed lists of atoms and hence no new covalent bonds can be formed during the simulation. Bonded interactions can be described by bond stretching, angle bending, and torsions or dihedrals. The potential energy function of bonded interactions may be written as

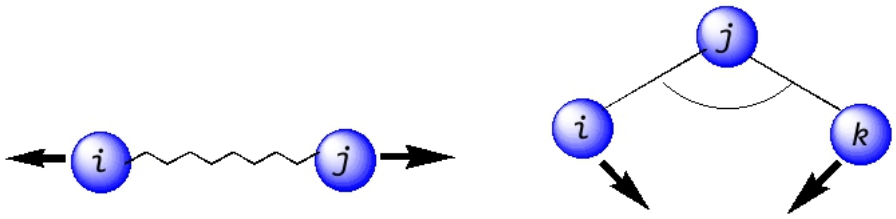
$$V_{bonded} = V_{bonds} + V_{angles} + V_{dihedrals}. \tag{3.2}$$

The bond stretching is a 2-body interaction determined between two covalently bonded atoms i and j at a distance r_{ij} from each other. For the stretching see Figure 3.2(a). It is often represented by harmonic potential or Hooke's law formula,

$$V_{bonds} = \frac{1}{2} k_r (r_{ij} - r_0)^2, \tag{3.3}$$

where k_r represents the force constant, and r_0 is the reference bond length value defined by the force field. Hooke's law is usually a sufficient and computationally efficient formula and thus widely used in different force fields. A more accurate representation, when needed, can be achieved by Morse potential.

The angle bending is a 3-body interaction defined between three atoms i, j, and k. For bending, see Figure 3.2(b). It is also often represented by a harmonic potential



(a) Bond stretching (b) Angle bending

Figure 3.2 *Bond stretching and angle bending. Their potential energy functions are often formulated by harmonic potentials.*

and thus equals

$$V_{angles} = \frac{1}{2} k_\theta (\theta_{ijk} - \theta_0)^2, \tag{3.4}$$

where k_θ is the angular force constant, θ_{ijk} is the angle between ij and jk, and θ_0 is the reference angle value defined by the force field. A more accurate form for angle bending can be derived by adding higher-order terms to the harmonic potential.

The terms of bond stretching and angle bending are often regarded as "hard" degrees of freedom, as considerable energies are required to cause significant deformations from their reference values. Instead, most of the structural variations and relative energies arise from a more complex interplay, that is, non-bonded interactions and dihedrals [76]. The non-bonded contribution is not discussed until the next section. The dihedral potentials concern a quartet of atoms and are thus 4-body interactions. There are two kinds of interactions, proper and improper. Thus, one may write

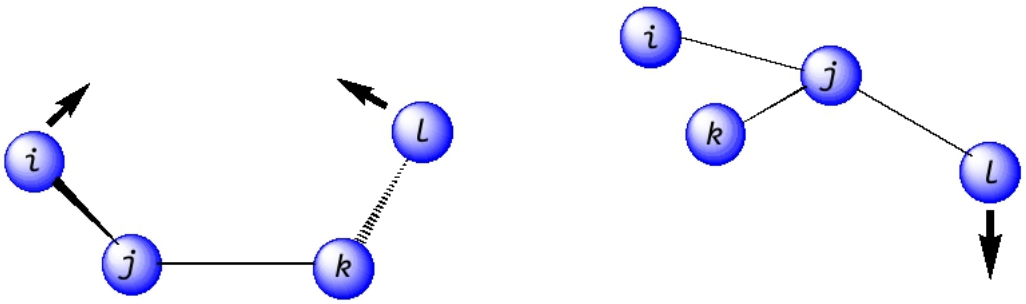
$$V_{dihedrals} = V_{proper} + V_{improper}. \tag{3.5}$$

The proper dihedrals are used to prevent bond rotations of atoms i, j, k and l. For rotation, see Figure 3.3(a). The proper dihedral potentials are commonly described either by periodic potential functions or potential functions of a cosine series expansion. The latter, the so called Ryckaert-Belleman potential, is also used in this study and equals

$$V_{proper} = \sum_{n=0}^{5} C_n (\cos\phi)^n. \tag{3.6}$$

In this form the constant C_n is defined in the force field and ϕ is the torsion angle.

Finally, the improper dihedrals are used to restrain chiral and planar centers. For instance, the aromatic ring of benzene is kept planar, and phosphate is held tetrahedral. For an example of improper dihedral, see Figure 3.3(b). The simplest improper



(a) Proper dihedral (b) Improper dihedral

Figure 3.3 *Examples of dihedral angles. (a) Proper dihedrals restrain the rotation of the bond between atoms j and k. (b) Improper dihedrals restrain out-of-plain bending (shown here), and molecules from flipping over to their mirror images (not shown here).*

dihedral representation is the harmonic potential form,

$$V_{improper} = \frac{1}{2} k_\xi (\xi_{ijkl} - \xi_0)^2, \tag{3.7}$$

where k_ξ is the force constant, ξ_{ijkl} represents the improper dihedral angle, and ξ_0 is its reference value defined by the force field.

3.2.2 Non-bonded Interactions

Non-bonded interactions describe the interplay between atoms that are not covalently linked to each other but remain within a certain distance. Unlike bonded interactions they are based on varying lists of atoms that are continuously updated during the simulation. Non-bonded interactions can be described by van der Waals interaction, that is, short-range repulsion and long-range attraction between two atoms. For this there are two commonly used potentials, the Lennard-Jones (LJ) potential and the Buckingham potential. Both potential functions consist of two terms, the repulsion and the attraction term. Additionally, if interacting atoms are charged, a Coulomb potential term has to be included in the non-bonded interactions. Hence, the potential energy function of non-bonded interactions is written

$$V_{non-bonded} = \underbrace{V_{repulsion} + V_{attraction}}_{van\ der\ Waals} + \underbrace{V_{Coulomb}}_{electrostatics}. \tag{3.8}$$

The Buckingham potential has a more flexible and realistic repulsive term than Lennard-Jones, but it is more expensive to compute [75]. The Lennard-Jones poten-

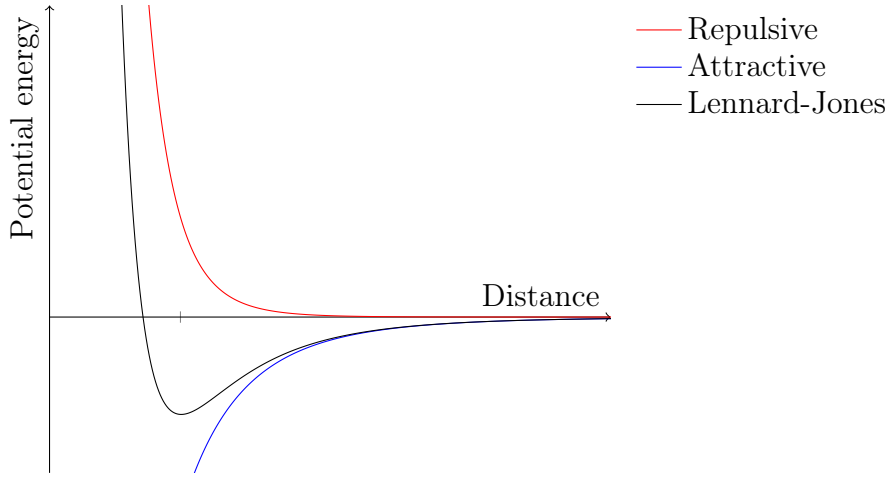


Figure 3.4 *Schematic Lennard-Jones potential energy function that describes the interaction between non-bonded particles. The potential comprises a repulsive term of the form* $\frac{1}{r^{12}}$ *and an attractive term of the form* $-\frac{1}{r^6}$.

tial V_{LJ} between the atoms i and j at a distance r_{ij} from each other equals

$$V_{LJ} = V_{repulsion} + V_{attraction} = \frac{{C_1}^{12}}{r_{ij}^{12}} - \frac{{C_2}^{6}}{r_{ij}^{6}}, \tag{3.9}$$

where the parameters C_1 and C_2 depend on atom types. A schematic plot of the Lennard-Jones potential is shown in Figure 3.4.

The Coulomb potential $V_{Coulomb}$ between two charged particles is given by

$$V_{Coulomb} = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon_r r_{ij}}, \tag{3.10}$$

where q_i and q_j represent the charges of the particles, ϵ_0 equals the vacuum permittivity, and ϵ_r is the relative permittivity.

3.3 Temperature and Pressure Are Set Analogically to Experiments

Next in the work flow, it is time to consider ideal laboratory conditions for the studied systems and how it is possible to mimic them computationally. MD simulations are occasionally performed in microcanonical conditions, where the number of particles N, and the volume V, and energy E of the system are held constant (NVE conditions). However, in experiments a certain temperature T and pressure p are crucial for specific biological and chemical processes. The most common implementations in MD are therefore constant volume and temperature (NVT) conditions,

and constant pressure and temperature (NpT) conditions. In this study the latter, constant temperature and pressure ensemble was chosen.

Common methods for maintaining constant temperature in the simulation are the weak coupling scheme of Berendsen [79], the extended ensemble Nosé-Hoover scheme [80, 81], and the velocity rescaling (v-rescale) scheme [82]. In the Berendsen algorithm the system is coupled to an external heat bath with a given temperature T_0. The algorithm corrects the deviation of the temperature from this value according to

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau}, \tag{3.11}$$

where τ is a time constant. Thus, the temperature deviation decays exponentially, and the strength of the coupling can be varied to the requirements with the time constant. This is an advantage of the Berendsen method. However, the method suppresses the fluctuations of the kinetic energy making the sampling incorrect. The error has been corrected in the velocity rescaling method used in this study. It is essentially a Berendsen thermostat, but additionally it has a stochastic term that ensures a correct energy distribution. The term equals

$$dK = (K_0 - K)\frac{dt}{\tau_T} + 2\sqrt{\frac{KK_0}{N_f}}\frac{dW}{\sqrt{\tau_T}}. \tag{3.12}$$

In the form K is the kinetic energy and K_0 is its reference value. N_f is the number of degrees of freedom and dW a Wiener process. The parameter τ_T is close to the time constant τ.

Common methods for simulating constant pressure are Berendsen algorithm [79], the extended ensemble Parrinello-Rahman approach [83], and the velocity Verlet variants, the Martyna-Tuckerman-Tobias-Klein (MTTK) implementations of pressure control [84]. In this study the Parrinello-Rahman coupling was implemented. In this barostat, the box vectors represented by the matrix $\mathbf{b}$ obey the matrix equation of motion,

$$\frac{d^2\mathbf{b}}{dt^2} = V\mathbf{W}^{-1}\mathbf{b}'^{-1}\mathbf{P} - \mathbf{P}_{ref}. \tag{3.13}$$

In this form, the volume of the box is denoted by V, and $\mathbf{W}$ is a matrix parameter defining the strength of the coupling. $\mathbf{P}$ is the pressure and $\mathbf{P}_{ref}$ the reference pressure matrix. The equations of motion for the particles are also changed. Thus the Parrinello-Rahman modification becomes

$$\frac{d^2\mathbf{r}_i}{dt^2} = \frac{\mathbf{F}_i}{m_i} - M\frac{dr_i}{dt}, \tag{3.14}$$

$$M = \mathbf{b}^{-1}\left[\mathbf{b}\frac{d\mathbf{b}'}{dt} + \frac{cd}{dt}\mathbf{b}'\right]\mathbf{b}'^{-1}. \tag{3.15}$$

Here $\mathbf{W}^{-1}$ is the inverse mass parameter matrix that determines the strength of coupling and how the box can be deformed. For its determination one needs to provide the approximate isothermal compressibilities β and the pressure time constant τ_P. Thus a matrix element becomes

$$\left(W^{-1}\right)_{ij} = \frac{4\pi^2\beta_{ij}}{3\tau_P^2 L}, \tag{3.16}$$

where L equals the largest box matrix element. The Parrinello-Rahman algorithm can be combined with any of the temperature coupling methods available in GROMACS and thus its usage together with the velocity rescaling scheme in this study is justified.

3.4 Periodic Boundaries Make System Infinite

As mentioned before, even though the studied system is originally finite, there are ways to make it infinite. The need arises from the boundaries of the system, where the matter would otherwise interact with a container or vacuum. At these interfaces tedious edge effects might come up, and additionally the original system should be rather big in the first place. That is why periodic boundary conditions (PBCs) are often used in molecular dynamics simulations. Using PBCs aims at minimizing edge effects. The idea is to have one *unit cell*, and to surround the cell by translated copies of itself called *images*, see Figure 3.5. In practice, when a protein or any molecule in the system enters a wall during the simulation, the entered part appears on the other side of the box. In this way, the originally finite system becomes infinite and ideally there are no more disturbing edge effects. However, artifacts may also occur due to the periodic boundary conditions. Especially crucial is the size of the unit cell. A small system is often favored to make the computation faster, but in small systems, particles might experience the same interactions more than once due to the PBCs, which may ruin the simulation.

For dealing with the interactions within the lattices of unit cells, different kinds of algorithms are available. For short non-bonded interactions the periodic boundary conditions in GROMACS are used in combination with a *minimum image convention* meaning that only one, the nearest, image of each particle is considered. For long-range non-bonded interactions this might not be accurate enough, and instead lattice sum methods such as the Ewald sum and the particle mesh Ewald (PME) technique, are used. In the PME used in this study the charges are assigned to a

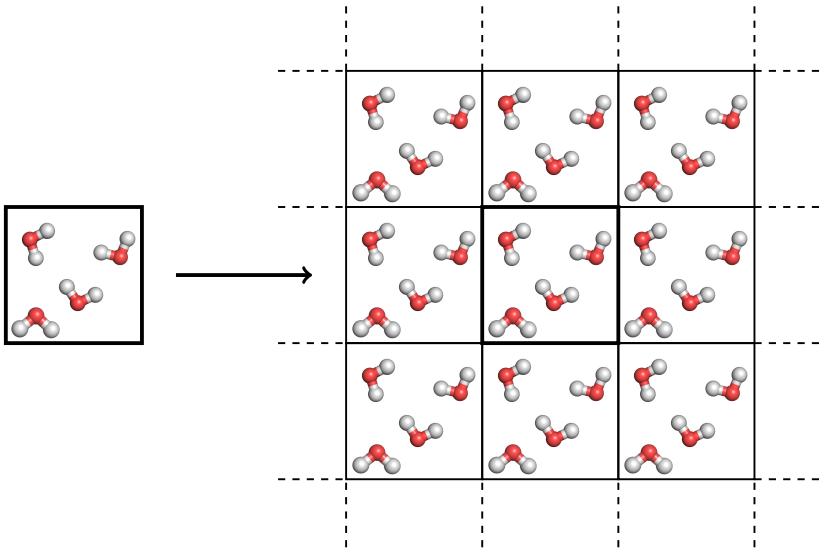


Figure 3.5 *Implementation of periodic boundaries for a 2-dimensional box with four water molecules. The picture demonstrates how a finite unit cell (on the left) is used to describe an infinite lattice (on the right) when the periodic boundary conditions are implemented. The picture is from Reference [85].*

grid using interpolation. The implementation in GROMACS uses cardinal β-spline interpolation, which is referred to as smooth particle mesh Ewald (SPME) [86].

3.5 Molecular Dynamics Is Based on Newtonian Mechanics

Before running a simulation, the system has to be energy minimized. Especially, if the starting structure is far from its equilibrium, the forces may be too large, which in turn might result in simulation failure. Additionally, energy minimization reduces the thermal noise in the structures and potential energies so that they can be compared better. Finally, the simulation may be started. First, it is good to know that molecular dynamics relies on Newtonian mechanics, and thus quantum mechanics is not considered. What this means is that the electronic motions in the system are ignored, and instead the motion of atoms is only described by their nuclei [76]. The simplification is denoted as the *Born-Oppenheimer* approximation. In this way calculations on systems containing significant numbers of atoms can be performed at a reasonable calculation expense. For most of the atoms, the approximations work well at normal temperatures [75]. However, quantum mechanical simulation methods are also available, when needed [76].

A simulation is ran based on simulation parameters defined by the user. The parameters give all instructions to the simulation program concerning the performance, what is the total length and time step, and what information is saved and how often,

for example. Additionally, all options regarding the simulation, electrostatics, van der Waals interaction, pressure, temperature, periodic boundaries, and others are set in the simulation parameter file. In the simulation, the same calculation protocol is repeated again and again. Time step is the time interval that defines how frequently the calculations are repeated. At each step the following calculations are performed. First, the forces on atoms are derived from the potential energy functions. The force acting on the atom i at position $\mathbf{r}_i$ may be written as

$$\mathbf{F}_i = -\frac{\partial V_i}{\partial \mathbf{r}_i}, \tag{3.17}$$

where V_i represents the potential energy. Secondly, the corresponding Newton's equations of motions are solved. For the atom i having a mass m_i this equals

$$\mathbf{F}_i = m_i \frac{d^2\mathbf{r}_i}{dt^2}. \tag{3.18}$$

These two derivations are carried out simultaneously. There are many algorithms for integrating the equations of motion. However, all algorithms assume that positions and dynamics properties, such as velocities and accelerations, can be approximated by Taylor series expansions. One of the most widely used method is the *Verlet algorithm* [87]. This algorithm uses the position and acceleration at time t, and the position from previous step at time $t - \delta t$. Hence, the Taylor series expansions may be written as functions of the positions $\mathbf{r}(t)$, and $\mathbf{r}(t-\delta t)$, and the acceleration $\mathbf{a}(t)$. The velocities do not explicitly appear in the Verlet algorithm, which is one of the disadvantages of the method. One of the variations of the Verlet algorithm that does explicitly include the velocities is the *leap-frog* algorithm [88]. This method uses the following relations for the position $\mathbf{r}$ and the velocity $\mathbf{v}$:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t + \frac{1}{2}\delta t), \tag{3.19}$$

$$\mathbf{v}(t+\frac{1}{2}\delta t) = \mathbf{v}(t - \frac{1}{2}\delta t) + \delta t \mathbf{a}(t). \tag{3.20}$$

However, one disadvantage of leap-frog is that positions and the velocities are not synchronized which means that the contribution of kinetic energy to the total energy can not be calculated simultaneously when the positions are defined. Nonetheless, both the Verlet and leap-frog algorithms are straightforward and have modest storage requirements.

After solving forces from potential energy functions and Newton's equations the overall configuration of the system can be updated. Additionally, output data may be written down. The output file describing the time evolution of a quantity, such as

atomic coordinates, is termed a *trajectory* file. Usually it is not worth saving the quantities to the trajectory at each time step to limit the usage of memory capacity. Too dense data is not always even relevant for analyses that are carried out as the final task. Many simulation packages offer ready tools for analysis. What exactly is analyzed depends on the studied system. If the dynamics of a protein are of interest, for instance, a common analysis is root-mean-square deviation (RMSD) that measures the average deviation of atoms from a reference structure. This enables comparing the simulation structure to experimentally resolved structures. After all, an experimental structure is just a snapshot and might have artifacts due to deleted loops and binding of antibodies, for example, that actually might change the structure.

3.6 Limitations

Molecular dynamics has a rather simple idea, as described above. However, the method is extremely delicate. It may produce brilliant results with ultimate atomistic resolution but like any other method it also suffers from limitations. This is due to the many approximations and assumptions needed to make the computations efficient, and due to the parametrization that is based on experiments and quantum mechanical calculations. First, the simulations are classical, meaning that the atomic motion is described by Newtonian mechanics. For most of the atoms at normal temperatures this is all right but there are few exceptions. Hydrogen atom possessing one single proton may experience quantum phenomena such as tunneling. Such processes can not be dealt with classical mechanics. Additionally, in the Born-Oppenheimer approximation electrons are expected to remain in their ground state. Thus, electron transfer processes and electronically excited states can not be treated. Neither can chemical reactions be considered without special algorithms.

Second, the role of the used force field can not be overvalued. If the description of the underlying potential energy functions and parameters fails, nothing good can come out of it. For this there is a saying in computer sciences that hears "garbage in, garbage out". However, the force fields are constantly updated to become more accurate, comparable to experimental results, and comparisons of different force fields are also carried out once in a while [78, 89]. This helps one to choose the most accurate force field for different purposes. Additionally, although the force fields are not really parts of the simulation method, the forms of the forces that can be used in a particular program is limited. The GROMACS force field for example is pair-additive (apart from long-range Coulomb forces), meaning that all non-bonded forces result from the sum of non-bonded pair interactions. As a result it can not

incorporate polarizabilities and does not contain fine-tuning of bonded interactions, which give rise to some limitations.

The third aspect to the limitations of molecular dynamics is about what rationally can be performed. In principle, the method offers unlimited temporal and spatial resolution, a feature which often restricts experimental procedures. However, the limitations rather come along with bigger scales, that is how long and big simulations are possible to carry out and what can be deduced from them. Today, the temporal time scales of atomistic simulations span some microseconds. The number of atoms in such a system may roughly be a million. These scales are only reached with the aid of massively parallel supercomputers. To account for reproducibility, MD simulations often need to be repeated a few times to get reliable results which make it even more time and computational resource consuming. Finally, the current achievable scales in MD are still at the limit where it is not clear whether protein functions, as a case in point, can be studied. Simply put, a simulation must be able to span similar time scales as the studied real-life processes do.

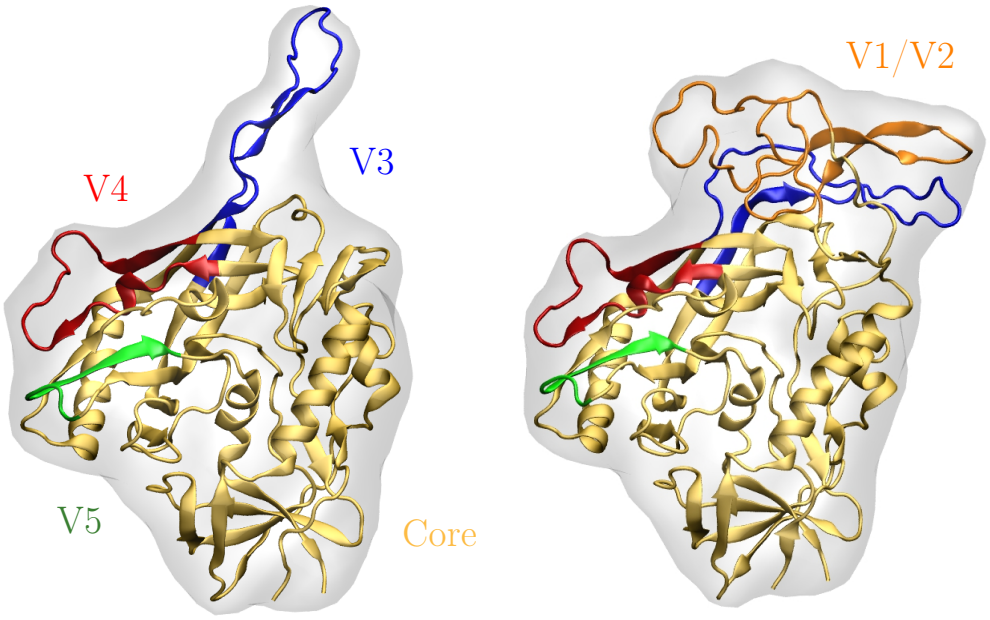
4. SIMULATION SYSTEMS AND PARAMETERS

In this Chapter, the preparation of the studied systems and their simulation details are presented. Neither of the tasks were carried out in the framework of this Thesis. Instead, the performance is all thanks to the professional co-workers of the project, Dr. Pekka Postila (systems), and Dr. Moutusi Manna (systems and simulations). However, presenting their valuable work is necessary and of importance for the analysis part that is the core of this Thesis. In the building process it is extremely crucial to pay attention to the underlying biology, and thus special emphasis was laid on the validity of choices. Good sources for gaining such information are the Protein Data Bank (PDB) archive, the single worldwide repository of information about the 3-dimensional structures of proteins and nucleic acids, as well as the Universal Protein Resource (UniProt), a comprehensive resource for protein sequence and annotation data. Additionally, HIV databases funded by the Division of AIDS of the National Institute of Allergy and Infectious Diseases (NIAID), a part of the National Institutes of Health (NIH), is a useful source of information. The computing facilities for this study were provided by the Partnership for Advanced Computing in Europe (PRACE) DECI-10 project HIV1-GSL and the Finnish IT Center for Science (CSC).

4.1 Models of Monomer and Trimer

As pointed out in Section 2.1, HIV-1 has a variety of subtypes that further are divided into isolates. In this study, the major group M and its subtype B were of interest. Most of different kinds of functional studies so far have used the isolate designated HXB2. However, this is not a native isolate, but an isolate generated in laboratory as a recombinant virus. For this study, the second most studied and a *native* isolate, extracted from humans with HIV-1 infection, was chosen. The isolate is designated YU-2. In comparison to the HXB2 isolate, YU-2 is an even more difficult target for antibodies. In the UniProt database the isolate used can be found with the identifier P35961.

Many molecular dynamics simulations quite similar to this study have been perfor-

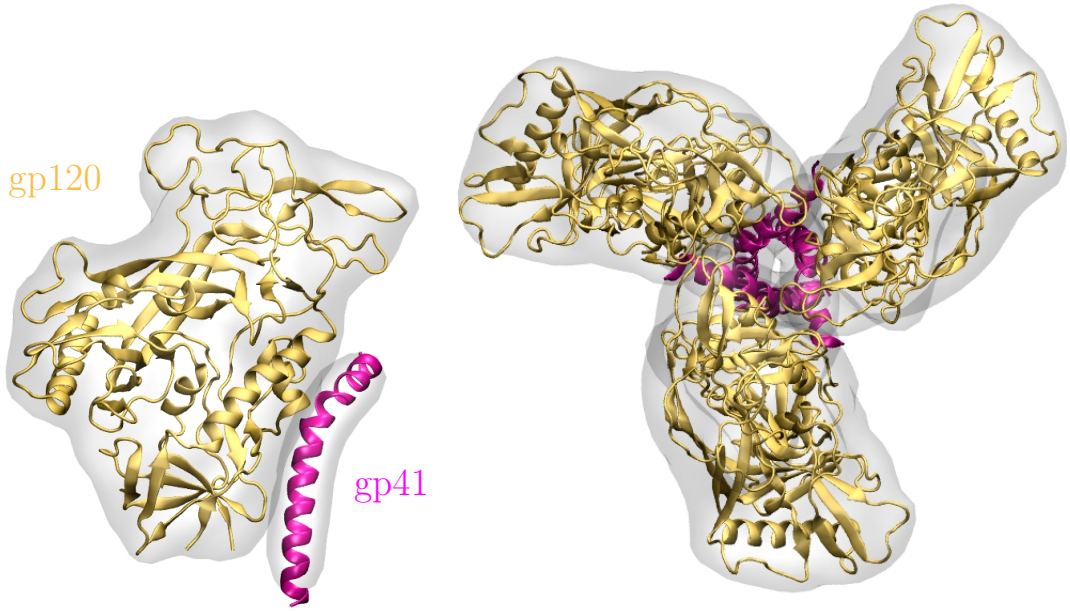


(a) *Truncated* monomer (b) *Complete* monomer

Figure 4.1 *Ribbon representations of the gp120 monomers built for this study. (a) The truncated monomer comprised of the gp120 core and three variable loops (V3, V4, and V5). (b) The complete monomer comprised of the gp120 core and all five variable loops (V1, V2, V3, V4, and V5).*

med earlier to study gp120 dynamics. However, these studies have only considered the core and the variable loops V3, V4, and V5 [5, 6, 7, 90, 91, 92]. The variable loops V1 and V2 have been absent without exception. However, as described in Section 2.2.3, the major variable loops, V1 to V3, play essential roles in gp120 functions, and thus they should all be taken into account to get the whole picture. What is more, the V1/V2 complex comprises a long sequence that lies in the immediate vicinity of the V3 loop. Thus, the three major loops might well function in concert. Additionally, the trimeric Env has not been studied with the aid of molecular dynamics simulations either. Yet, it is known that understanding gp120 dynamics in its native conformation, in the trimer, is crucial for antibody recognition in vivo. In this study, all variable loops as well as the trimer have been taken into account.

To fulfill these aims two gp120 monomers and a trimer were built. To this end, three structures from the PDB archive were used. Their PDB identifiers are 4NCO [33], 1G9N [93], and 4JZZ [94]. The resolution of these structures is 4.70 Å, 2.90 Å, and 1.49 Å, respectively. As the preparation of the model systems was not part of this Thesis, the detailed description of the process is not presented here. There are three gp120 systems in total. The first one is a gp120 monomer possessing only three varible loops, V3 to V5, and is termed the *truncated* monomer. The second one is a gp120 monomer possessing all five variable loops, V1 to V5, and is termed



(a) Subunit of trimer (b) Trimer

Figure 4.2 *Ribbon representation of the gp120 trimer built for this study. (a) The subunits of the trimer are composed of noncovalently linked gp120 and gp41. The gp120 subunits are all identical to the complete monomer. (b) The trimer is composed of three gp120-gp41 subunits. The subunits are termed 1st, 2nd, and 3rd unit in this study.*

the *complete* monomer. The monomer structures are shown in Figure 4.1. The third system is a trimer consisting of three gp120-gp41 units. Each gp120 unit is identical to the *complete* monomer. For simplicity, the units are termed 1st, 2nd, and 3rd unit. The trimer is shown in Figure 4.2. Until today, the most advanced molecular dynamics studies have only considered the gp120 core with its three variable loops, V3 to V5. Thus, in this study there is a great possibility to directly analyze effects of the V1/V2 domain inclusion by comparing the *truncated* and *complete* monomers. In addition, there is a chance to compare the gp120 dynamics in the monomer and trimer conformations. Finally, the systems are shortly described in Table 4.1.

However, it has to be mentioned that the so-called *complete* monomer is not fully inclusive, which is simply due to the lack of its complete structure. As pointed out in Section 2.2.2, the crystallization of gp120 has not been straightforward. Until today, neither the complete 3-dimensional structure for gp120 monomer nor for the trimer is known. By the time of the system preparation, the used PDB structures gave the most accurate achievable and biologically relevant model for gp120. However, towards the end of this Thesis project a new crystal structure came out that would have given a somewhat better description [57]. In this new structure the N- and C-terminus of gp120 are more extent. Additionally, gp41 is also more extent.

Table 4.1 *Simulation systems.*

Type	Name(s)	No. of gp41	No. of gp120	Variable loops	Length
Monomer	Truncated monomer	0	1	V3 to V5	1 μs
Monomer	Complete monomer	0	1	V1 to V5	1 μs
Trimer	1st, 2nd, 3rd unit	3	3	V1 to V5	1 μs

However, in this study gp120 was of interest and gp41 was only included to keep the trimer as whole. Those parts of gp41 needed to hold the trimer together (the heptad repeat in the trimer core) did already exist and were used here. For comparison, gp120 and gp41 used here were superimposed with the new crystal structure. The superimposition is shown in Figure A.1. All in all, the structures used in this study are still relevant considering the main targets of the study that were the major variable loops.

The *complete* monomer was comprised of 437 and the *truncated* of 367 residues. Both monomers were hydrated with about 40000 water molecules. The trimer was comprised of 1422 residues and was hydrated with about 130000 water molecules. Additionally, sodium and chlorine ions were added such that the concentration corresponded to the native physiological environment with an ion concentration of 0.15 mol/dm^3. For the parametrization of all molecules and ions, the OPLS-all atom force field [95] was used, and for water the TIP3P model was employed as it is compatible with OPLS parametrization [96]. Finally, the conserved and variable domains were defined with the aid of the UniProt database. The gp120 structure used in this study was simply aligned to the database structure with the identifier P35961. The resulting domain definitions are shown in Table 4.2. The first residue index of gp120 was 44 and the last was 480.

4.2 Simulations and Parameters

Three unbiased molecular dynamics simulations were performed in order to investigate the dynamics of gp120. For this the GROMACS 4.6.5 package [75] was used. Before the simulations, energy minimization calculations were carried out for each system to find the local potential energy minimum near the starting structure. For this, two algorithms were applied, the steepest descent and the conjugate gradient algorithms. After energy minimization, simulations spanning 1 microsecond were performed. For the numerical integration, a time step of 2 fs was set. In all three

Table 4.2 *The definition of the residue compositions of gp120 domains.*

Domain	Truncated	Complete	Length
C1	44–123	44–130	80/87
V1	-	131–155	0/25
V2	-	156–193	0/38
C2	194–292	194–292	99
V3	293–326	293–326	34
C3	327–380	327–380	54
V4	381–406	381–406	26
C4	407–448	407–448	42
V5	449–459	449–459	11
C5	460–480	460–480	21

directions the periodic boundary conditions with the usual minimum image convention were used. The LINCS algorithm [97] was used to preserve hydrogen covalent bond lengths. The simulations were run under NpT conditions. The reference temperature was set to 310 K. For temperature coupling, the v-rescale thermostat [82] with 0.1 ps time constant was used. Separate heat baths for the solvent and the solute were provided. The reference pressure was set at 1 bar. For pressure coupling, the Parrinello-Rahman barostat [83] with 1.0 ps time constant was used. Van der Waals interactions described by the Lennard-Jones potential were cut-off at 1.0 nm. Particle mesh Ewald method [86] with a real space cut-off of 1.0 nm, β-spline interpolation (order of 6), and direct sum tolerance of 10^{-6} was employed to describe electrostatic interactions.

5. RESULTS AND DISCUSSION

In this Chapter the analyzed results of the studied systems are presented. The systems were described in Table 4.1 and the domain definitions, essential for the analysis, were given in Table 4.2. Ready GROMACS and VMD tools were used for the analyses. Additional statistical calculations were carried out in Matlab. The pictures were created in VMD. The plots were created in Latex.

5.1 Loops Fluctuate More in Monomer than in Trimer

The stability of the three-dimensional structure of a globular protein can be measured by comparing the deviation of the structure during the simulation to a reference structure, such as, the starting structure. For this, the root-mean-square deviation (RMSD) can be used. Zero as value indicates that the structure is identical to the reference structure, whereas a large RMSD value means that the structure significantly deviates from the reference structure. Here, the GROMACS tool g_rms was used to measure RMSD as a function of time. The calculations were performed on the Cα atoms. First, the RMSDs from the monomer cores were calculated. Then, the variable loops were included in the calculation in pairs or one by one to see their contribution to the deviation in detail. The V4 and V5 loops were first included in the calculation, followed by the V3 loop and finally the V1 and V2 loops. All results are shown in Figure 5.1. The RMSDs from the monomer cores and from the V4 and V5 loops were very similar between the *truncated* and *complete* monomers, which suggests that these domains were mostly unaffected by the presence of the V1/V2 domain. When the V3 loop was included in the calculation, increase in the RMSD was noticed in both *truncated* and the *complete* monomer. The increase was somewhat larger in the *truncated* monomer. This suggests that the structure of the the V3 loop deviated more from its starting structure in the absence of the V1/V2 domain.

The RMSDs were accordingly calculated from the trimer units. All results are shown in Figure 5.2. The RMSDs from the cores and from the V4 and V5 loops were rather similar and of the same order than in the monomers. This suggests that the core

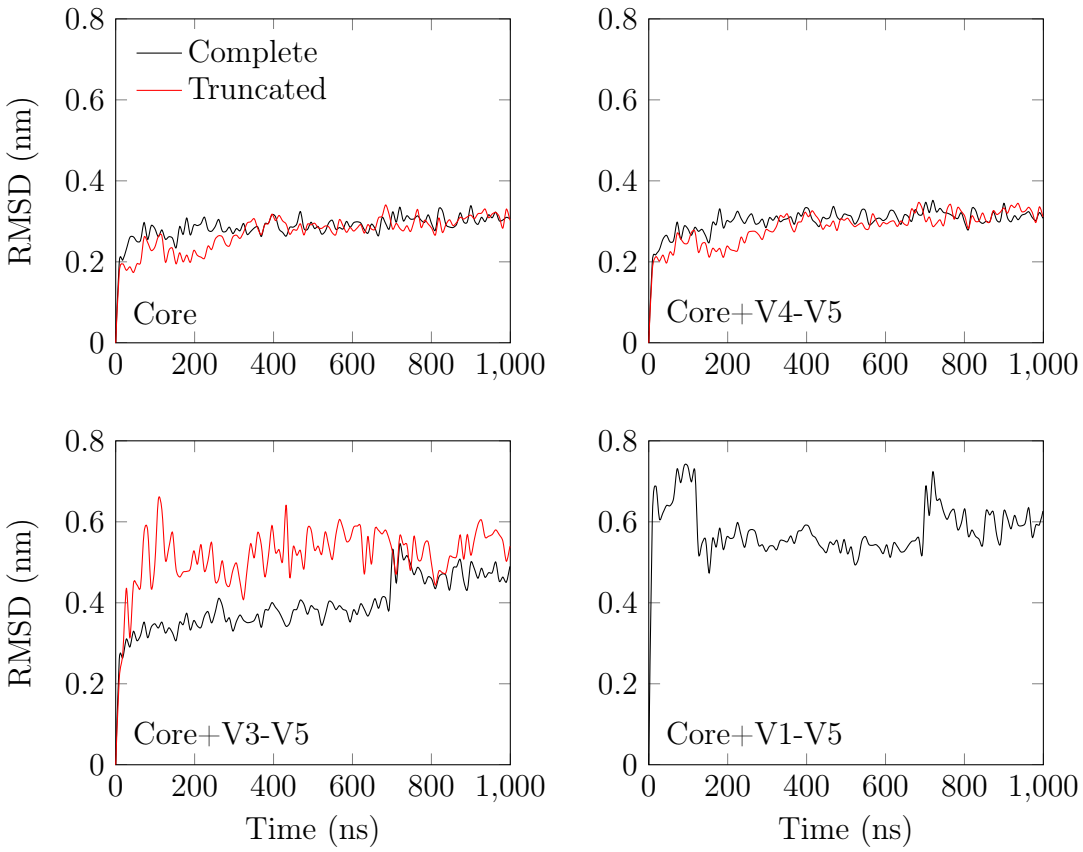


Figure 5.1 *The RMSDs from the truncated and complete monomers.*

structure and the V4 and V5 loops remain rather similar regardless of the V1/V2 domain and whether the context is a monomer or a trimer. The result is in agreement with the current understanding. When the V3 loop was included in the calculations, the RMSDs from the trimer units changed a bit differently: The RMSD from the *3rd* trimer unit increased significantly but the changes in the other two subunits of the trimer were rather minor. Finally, when the V1 and V2 loops were included in the calculations, the RMSDs increased in all gp120 subunits of the trimer. All in all, the differences between the trimer unit's RMSDs suggest that the major variable loops can play a bit different roles in different circumstances. This might be, for example, due to different interactions between the trimer units. For final comparison, the average RMSDs from each simulation were calculated with their standard deviations. The results are shown in Table B.1.

Next, the flexibility of the gp120 structure was studied. It can be measured by the magnitude of motion of atoms. For this, the root-mean-square fluctuation (RMSF) was used. In general, the larger the RMSF value of an atom, the more flexible it is. Here the GROMACS tool g_rmsf was used to measure average RMSF as a function of atom index. The calculation was performed on the Cα atoms. First,

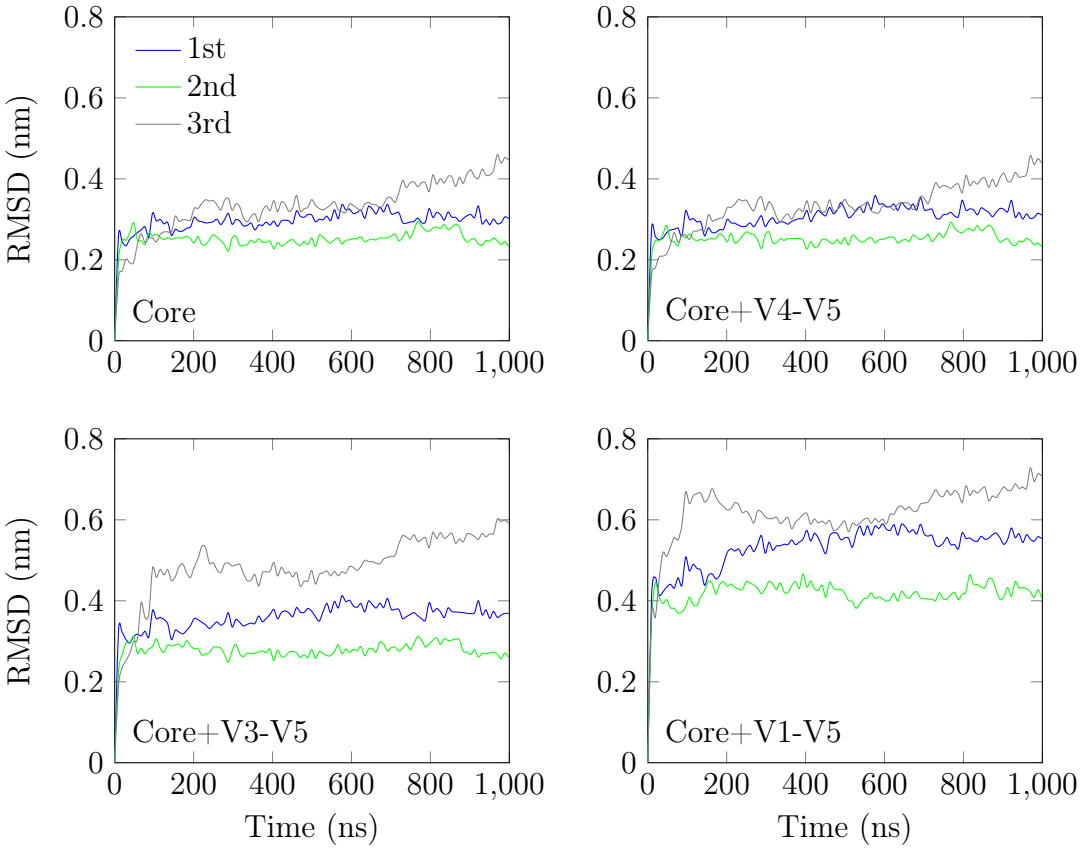


Figure 5.2 *The RMSDs from the gp120 subunits of the trimer.*

the RMSFs were calculated from the monomers. With the g_rmsf tool it was also possible to convert given RMSF values to β-factor values that were used to color a ribbon representation of the *truncated* monomer. Accordingly, the RMSFs from the *complete* monomer were calculated and the corresponding ribbon representation was drawn. The resulting graphs and pictures are shown in Figure 5.3. The RMSFs from the *truncated* monomer point out that the major variable loop V3 was the most flexible domain. The corresponding coloring in the ribbon representation shows the tip of the V3 loop in blue highlighting its flexibility. Most of the core is shown in red indicating its rigidity. The RMSFs from the *complete* monomer, in turn, show that the major variable loops V1, V2, and V3 were the most flexible domains. The corresponding ribbon representation similarly shows the variable loops V1 to V3 in green indicating their high flexibility. The results are in agreement with the current understanding.

The RMSFs were calculated from the gp120 subunits of the trimer. The ribbon representations of the subunits were colored similarly to the monomers. The graphs and pictures are shown in Figure 5.4. In general, one can notice that the RMSFs from the gp120 subunits of the trimer do not exhibit as high RMSF values as those

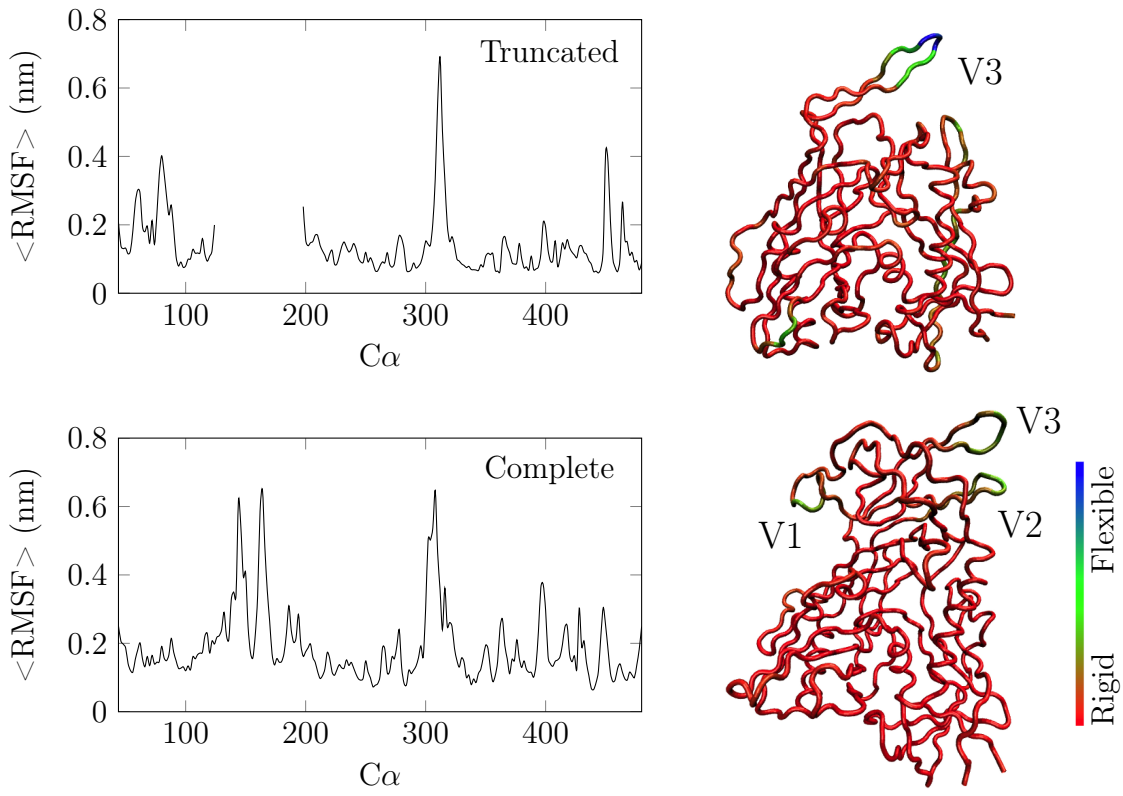


Figure 5.3 *The average RMSF from the monomer simulations and the corresponding ribbon representations. The first 200 ns of the simulations were excluded from the calculations.*

from the *complete* monomer. Especially, the peaks from the major variable loops seem to have decreased or vanished. Only the *2nd* trimer unit exhibits a high peak at the V1 loop. The ribbon representations similarly highlight the reduced flexibility of the major loops in all gp120 subunits of the trimer mostly showing these domains in red. As there has not been any similar simulations from the trimer so far, this is a new finding. Finally, all flexible residues from the variable loops of the monomers and trimer, arbitrarily defined by RMSF value greater than or equal to 0.40 nm, were listed in Table B.2.

5.2 Loop Tips Are More Mobile in Monomer than in Trimer

The mobility of the variable loops in the trimer context was studied and compared with the monomer systems. The size of globular proteins can be determined, for example, by the radius of gyration (R_G) that refers to the distribution of the components of an object around an axis. It defines the perpendicular distance from the axis of rotation to a point mass that gives an equivalent inertia to the original object. Here, the GROMACS tool g_gyrate was used to calculate the R_G from the

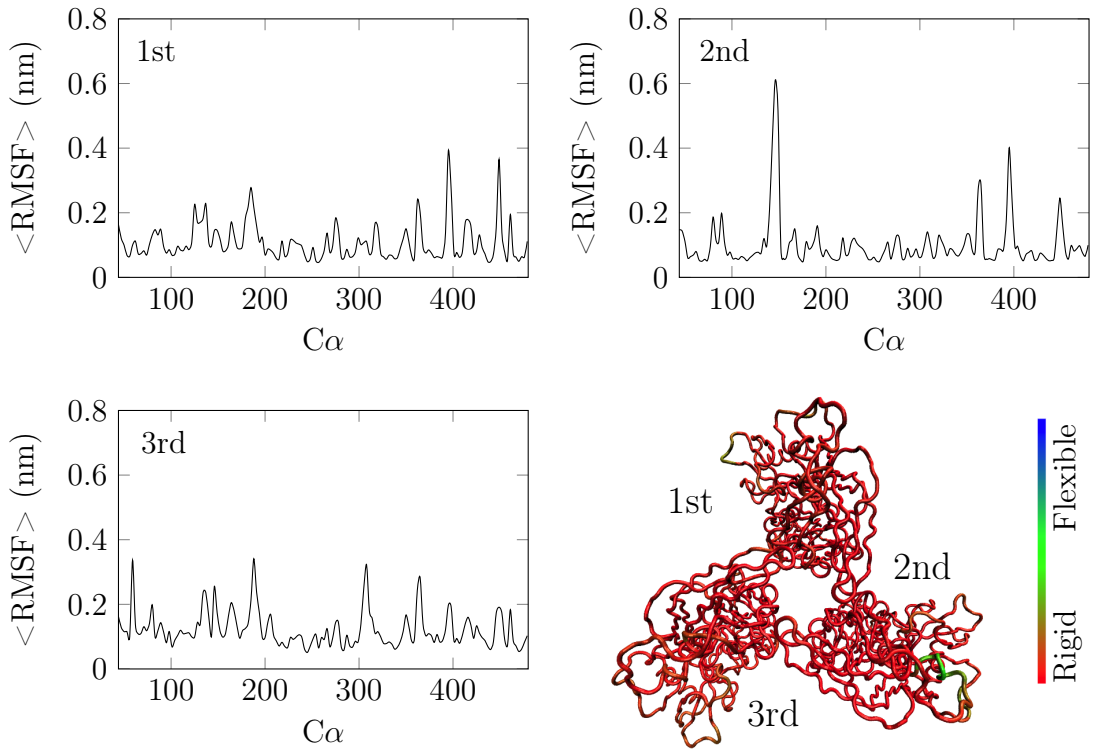


Figure 5.4 *The average RMSFs from the trimer simulation and the corresponding ribbon representation. The first 200 ns of the simulation was excluded from the calculations.*

Cα atoms as a function of time. First, the R_G from the cores of both monomers and the trimer units were calculated. Secondly, the V3 to V5 loops were included in the calculations, and finally, the V1 and V2 loops were included in the calculation. All results are shown in Table B.3. It was found out that the R_G from the *truncated* and *complete* monomers were very similar. Additionally, the R_G from the cores were almost the same in all systems. However, when the major variable loops were included in the calculations there were some differences between the trimer units. The R_G from the *1st* unit appeared to be systematically the smallest, and those from the *3rd* unit the greatest. The change in R_G between the units was not big, only about 8 %. However, it demonstrated that there might be some differences in the disposition of the variable loops during the simulation.

To study this more closely, center of mass (COM) distances between the gp120 core and the tips of the V1, V2, and V3 loops were calculated. For this, the GROMACS tool g_dist was used. Three residues from each variable loop tip were chosen for the calculation: in the V1 loop the residues 142 to 144, in the V2 loop the residues 164 to 166, and in the V3 loop the residues 308 to 310. First, the distance of the loops in the monomers were calculated. Then the distance of the loops in the trimer units were calculated. The results are shown in Figure 5.5.

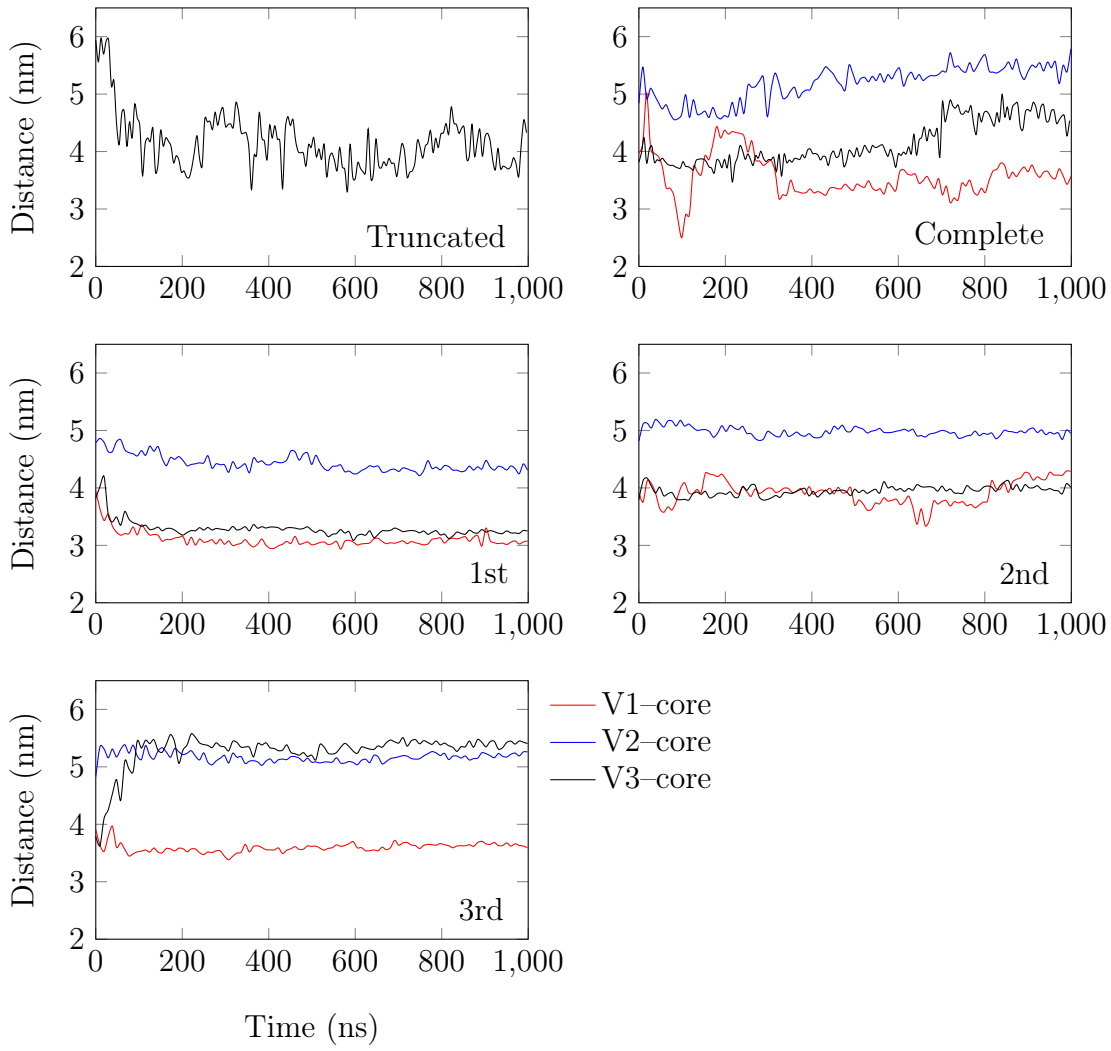


Figure 5.5 *The center of mass distance of the loop tips from the gp120 core.*

In general, the most drastic changes in the distances were seen in the monomers. In the *truncated* monomer the tip of the V3 tip shifted up to 2 nm and in the *complete* monomer the V1 tip up to 2.5 nm during the simulations. These most significant changes occurred in the beginning of the simulation. However, it can be clearly seen that in the trimer there was not similar loop tip oscillation than in the monomers. Instead, the loops seemed to stick to particular positions throughout the trimer simulation. What is more, all loop tip distances in the *1st* and the *2nd* trimer units were rather similar, whereas in the *3rd* trimer unit the distance of the V3 loop from the protein core was the largest of all systems. The results are in agreement with the R_G calculations which suggested that the *3rd* trimer unit exhibited the largest R_G. Based on this, it seems that the V3 loop of the *3rd* trimer unit is more exposed than the V3 loop in the *1st* and the *2nd* trimer units. In fact, it seems to be even more exposed than the V3 loop in the monomers.

Then, conformational distribution of gp120 was studied. First, the number of contacts between the major variable loops and the core as a function of time was calculated with the GROMACS tool g_mindist. Then, the number of the contacts was plotted as a function of R_G. Secondly, the COM loop tip distance from the gp120 core that was calculated earlier was similarly plotted as a function of R_G. The resulting distributions are shown in Figure 5.6. It was found out that the trimer units had rather dense conformational distributions in comparison to the monomers. The distributions of the monomers were more spread out. This is clearly seen in Figure 5.6(f), for example. The *3rd* trimer unit differed the most from the trimer units because it had the largest R_G which shifted the distribution to the right. The message is basically the same that was found out before: The V3 loop of the *3rd* trimer unit was the most exposed, and hence the *3rd* trimer unit also had the largest conformational distribution during the simulation. Additionally, it was found out that the distributions of the *complete* monomer were mostly concentrated on two spots. The difference was most clearly seen in Figure 5.6(e), where two spots are pointed out with arrows. The one distribution with smaller R_G resembles more the conformation of the *1st* and the *2nd* gp120 subunits, a more packed conformation. The other distribution with greater R_G is more similar to that of the *3rd* gp120 subunit, a less packed conformation. All in all, the monomeric gp120 seems to be more dynamic than the trimeric gp120. This is probably due to stabilizing interactions between the trimer units.

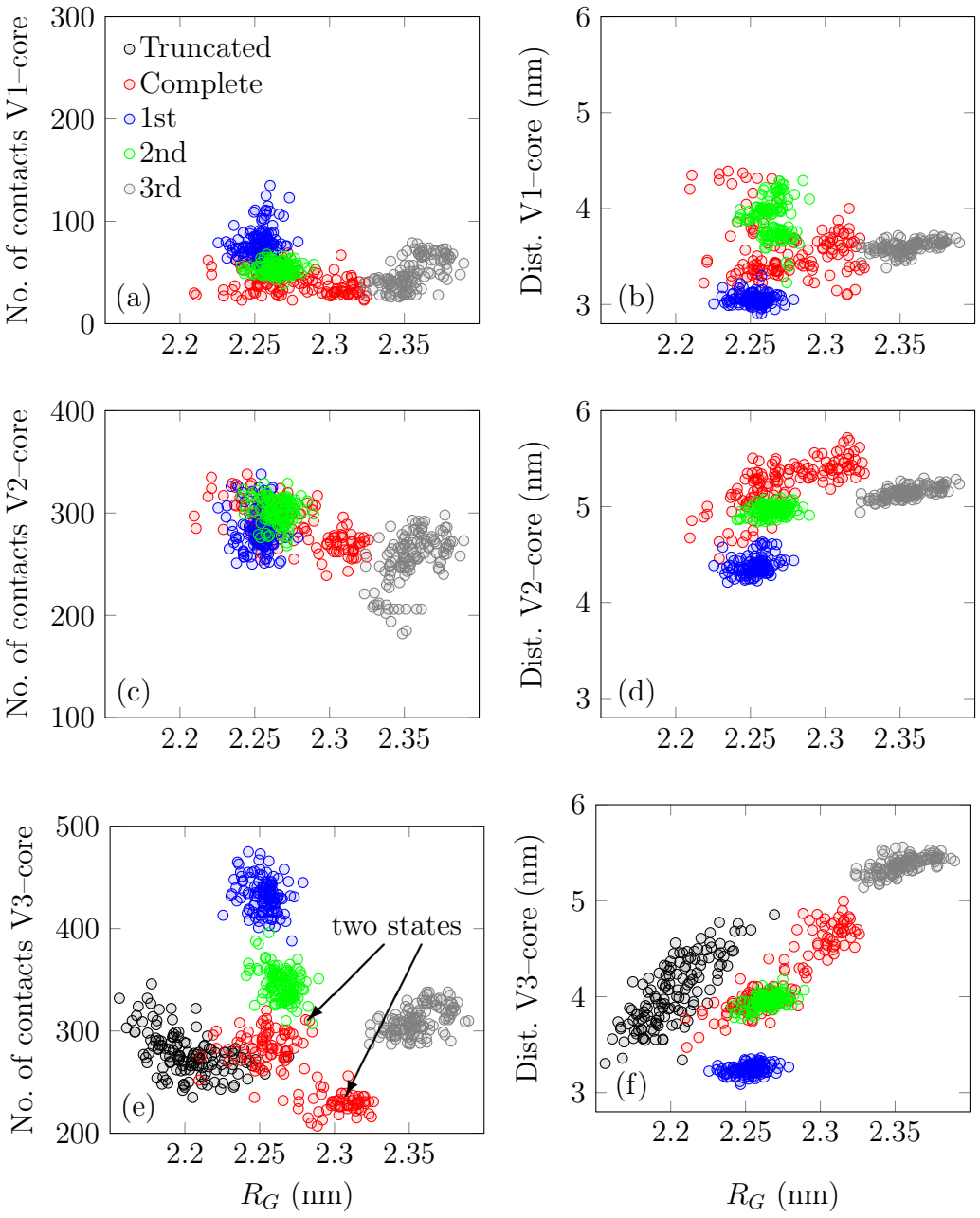


Figure 5.6 *Conformational distributions. On the left column: Distribution of the number of contacts between the loops and the core. On the right column: Distribution of the distance between the loop tips and the core. The radius of gyration was calculated without the V1/V2 domain to make the systems comparable. The first 200 ns of the simulations were excluded.*

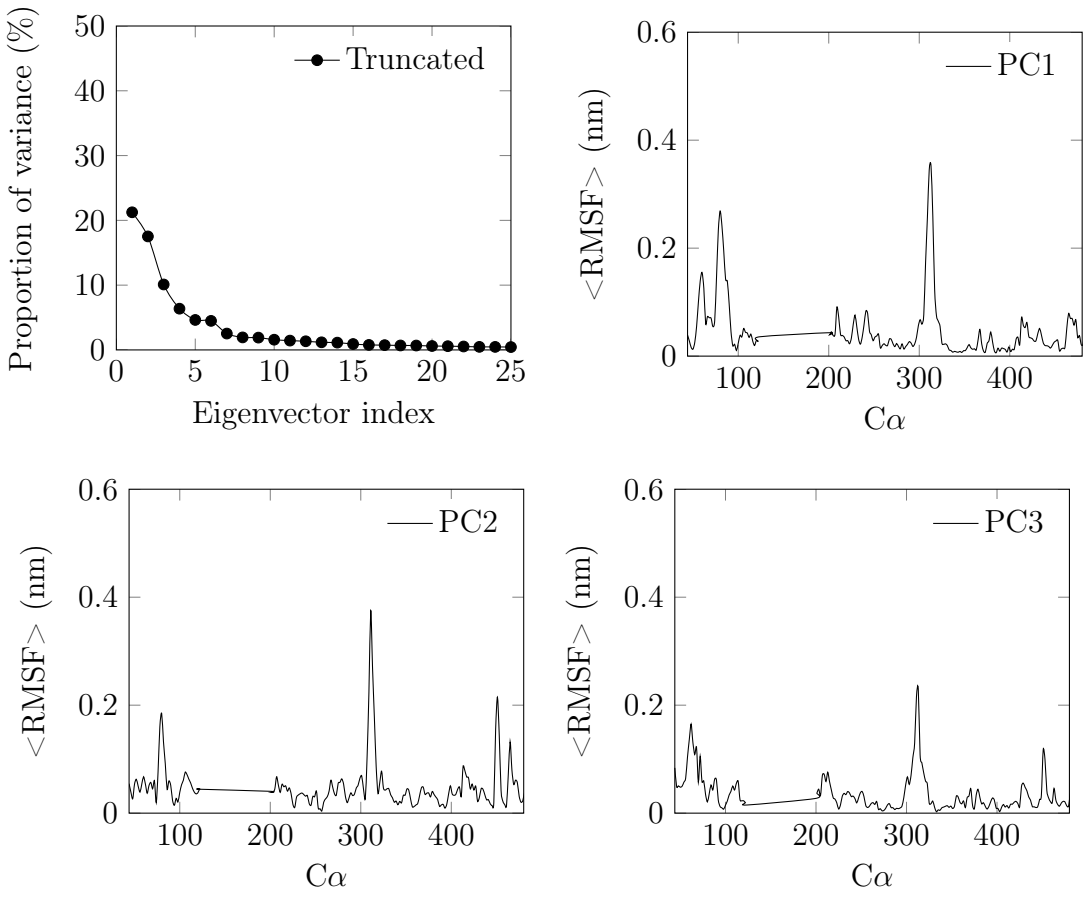


Figure 5.7 *The eigenvalues and the RMSFs from the principal components of the truncated monomer.*

5.3 Loops Show Concerted Motion in Monomer but not in Trimer

In order to gain a better understanding of the differences in the gp120 dynamics in different systems and to identify functionally relevant motion, principal component analyses (PCAs) were carried out. PCA is a statistical procedure where the principal components (PCs) of the data are looked for. PCs are the directions where there is the most variance in the data, that is, the directions where the data is the most spread out. In molecular dynamics it is often difficult to recognize the most relevant trends of the motion as everything moves at the same time. Hence, PCA is used to filter local (often fast) motion from collective (often slow) motion, the latter regarded as more relevant. Next, carrying out a PCA in the context of molecular dynamics is reviewed.

First in PCA of an MD trajectory, a covariance matrix is calculated. In the matrix an element in the *i,j* position tells the *covariance* between the i^{th} and j^{th} elements

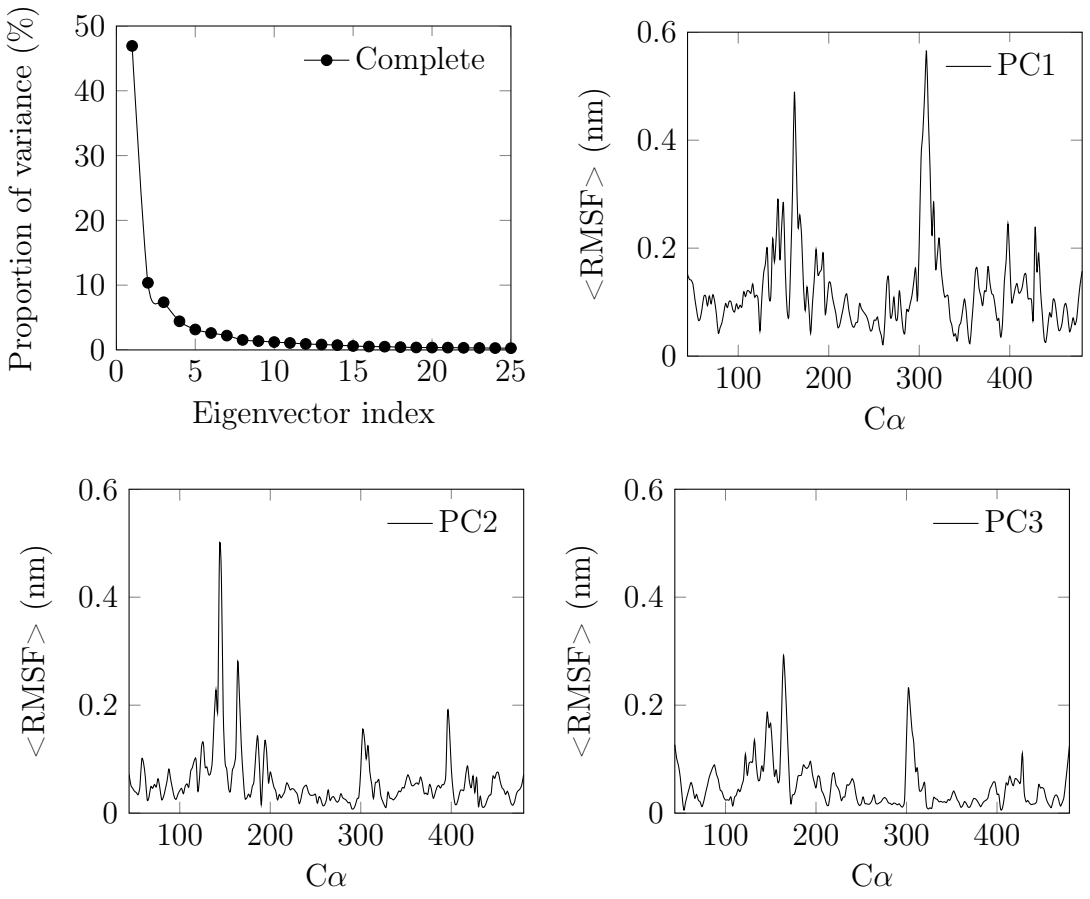


Figure 5.8 *The eigenvalues and the RMSFs from the principal components of the complete monomer.*

of a random coordinate. In these elements all directions x, y, z, have to be considered separately. For example, if two atoms 1 and 2 are described by the 3-dimensional coordinates (x_1,y_1,z_1) and (x_2,y_2,z_2) the covariance matrix will include the covariances from all possible coordinate combination, x_1x_1, x_1y_1, x_1z_1, x_1x_2, x_1y_2, and so on. For two atoms with 3-dimensional coordinates there will be $3^2 = 9$ combinations in total. The covariance between the coordinates of the atoms measures how much the coordinates change together. Similar behavior results in positive and dissimilar in negative covariance. After the calculation, the covariance matrix is further diagonalized. This is a useful operation as the *eigenvalues* and *eigenvectors* of a diagonalized matrix are known. In PCA, they turn out to be valuable objects.

Eigenvectors and eigenvalues exist in pairs. Eigenvectors are orthogonal vectors characterizing direction in the data. Each eigenvector has a corresponding eigenvalue that is a number telling how much variance there is in the data in that direction. Thus, the eigenvectors having the largest corresponding eigenvalues are the directions that possess the biggest variance in the data, which is exactly what was origi-

nally looked for. In fact, the eigenvectors with the largest eigenvalues are the principal components of the data. When principal components are found, the original x-y-z coordinate system can be forgotten and instead a more relevant representation can be gained by setting the principal components as new axes. Then, the original trajectory (or its frames) can be projected on the PCs instead of the familiar x, y, and z directions. In this way, the trajectory projected on the first principal component, for example, shows the motion in the direction of the greatest variance. The percentage of variance in each eigenvector direction is characterized by their corresponding eigenvalue.

Here, the GROMACS tool g_covar was used to calculate and diagonalize the covariance matrix from the trajectories. The tool produces a trajectory of the eigenvectors and lists the corresponding eigenvalues in increasing order. Typically, one easily notices that the magnitude of the eigenvalues decreases fast after the first three or so eigenvectors. Then, how many eigenvectors are chosen as principal components is one's own choice. Here, the principal components were chosen such that their motion described over 60 % of the total motion. The judgement was based on the magnitude of the eigenvalues. Then, the GROMACS tool g_anaeig was used to analyze the eigenvectors. The tool takes the original trajectory as input and projects its data on the chosen eigenvectors. This produces "filtered" trajectories that show the motion in the direction of each PC.

First, the PCs of the *truncated* monomer simulation were found. The first five eigenvectors were responsible for 60 % of the total motion of the system, and thus these eigenvectors were chosen as principal components. The original trajectory was filtered in regard to the PCs, and hence five new filtered trajectories were produced. Then, the RMSFs were calculated from the trajectories. The proportion of the variance of the the first 25 eigenvectors and the RMSFs corresponding to the first three PCs are shown in Figure 5.7. All RMSFs from the *truncated* monomer show a peak in the V3 domain. This suggests that the motion of the V3 loop plays a major role in the total motion in the *truncated* monomer.

Next, the PCs from the *complete* monomer simulation were looked for. The proportion of the first three eigenvectors corresponded to 64 % of the total motion, and hence these eigenvectors were chosen as PCs. Accordingly, three filtered trajectories were produced. The RMSFs from these trajectories were calculated. The proportion of the variance of the the first 25 eigenvectors and the RMSFs corresponding to the first three PCs are shown in Figure 5.8. The trajectory filtered in the direction of the first PC clearly shows two peaks, in the V1/V2 and V3 domains. Along the second and third PC the peaks decrease but are still present. This clearly suggest that the

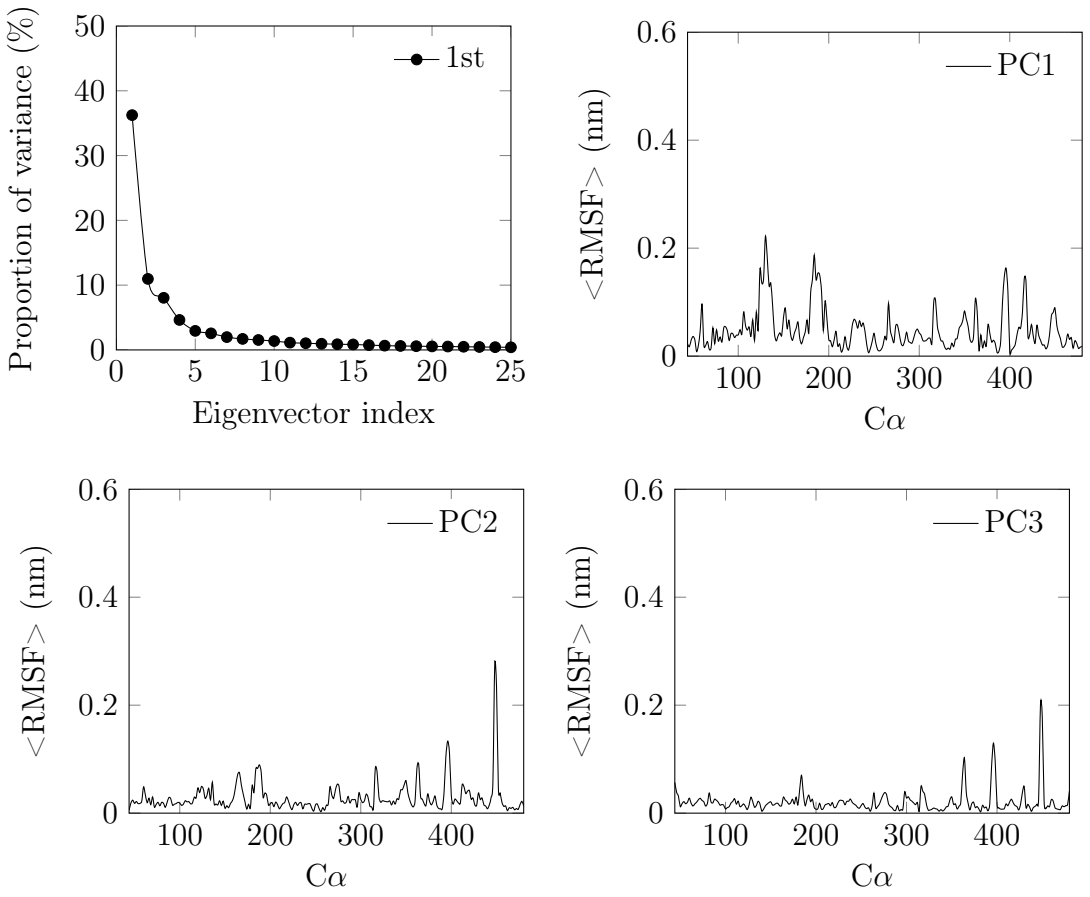


Figure 5.9 *The eigenvalues and the RMSFs from the principal components of the 1st gp120 subunit of the trimer.*

motion of the V1/V2 and V3 domains plays a significant role of the total motion and that the two domains move in concert in the *complete* monomer.

Then, the PCs from the trimer units were similarly looked for. In the *1st* trimer unit, first four eigenvectors corresponding to 60 % of the total motion were chosen as principal components. In the *2nd* unit, first five eigenvectors were chosen as PCs. They corresponded to 60 % of the total motion. Finally, in the *3rd* trimer unit, three first eigenvectors that corresponded to 64 % of the total motion were chosen as PCs. The proportion of the variance of the first 25 eigenvectors and the RMSFs from the first three filtered trajectories of the trimer units are shown in Figures 5.9 - 5.11. In comparison to the RMSFs from the filtered monomer trajectories, the RMSFs from the filtered trimer trajectories do not show as clear peaks in the V1/V2 and V3 domains. In fact, the peak of the V3 loop is only present in the *3rd* trimer unit. In the *2nd* trimer unit there is a high peak in the V1 domain but in the *1st* and the *3rd* trimer unit these peaks are clearly decreased. This suggests that the motions of the V1/V2 and V3 domains are restrained in the trimer. Also the concerted motion

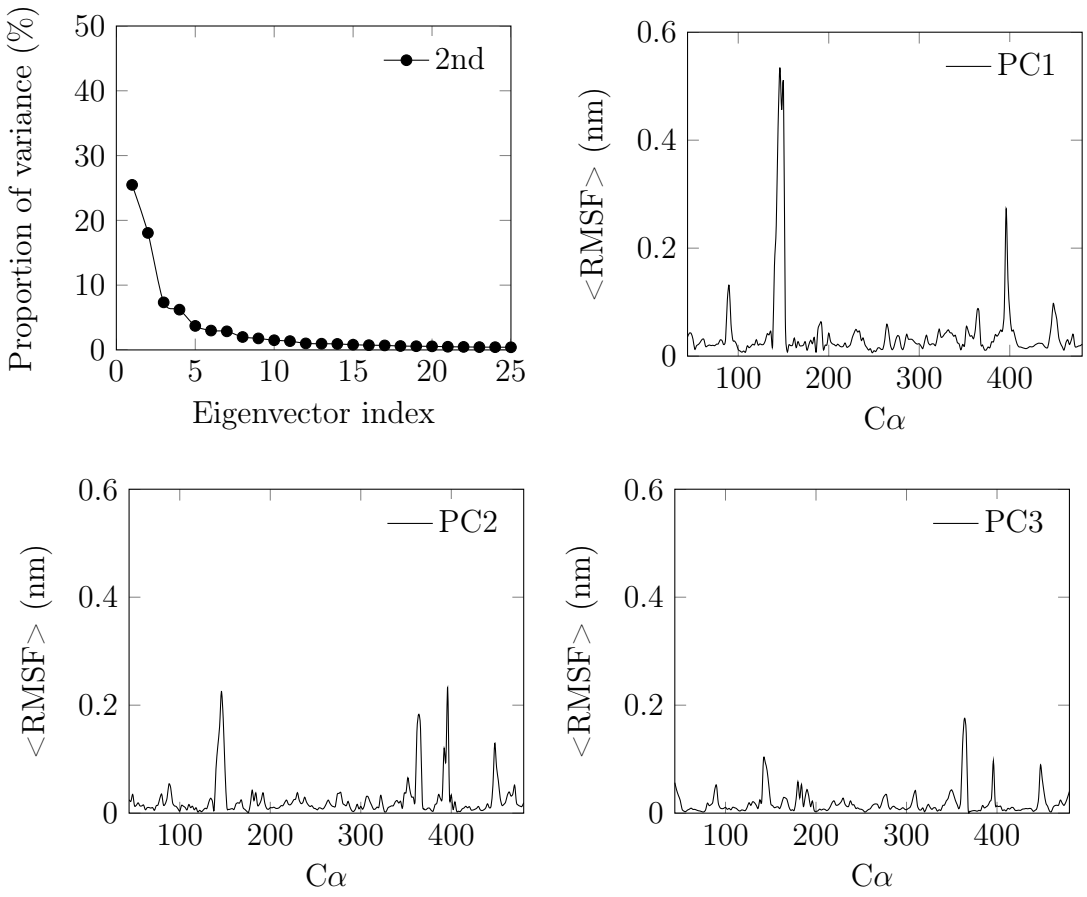


Figure 5.10 *The eigenvalues and the RMSFs from the principal components of the 2nd gp120 subunit of the trimer.*

of the two domains mostly seems to be lost.

Finally, the results of the principal component analysis was demonstrated by visualising the range of the movement of the V1, V2, and V3 loops along the first principal components. First, the range of the movement of the V3 loop in each system was visualized and is shown in Figures 5.12(b) - 5.12(f). The pictures clearly show that the range of the movement is the largest in the *truncated* and *complete* monomer. In the trimer units the range seems to be very small. Additionally, the V3 loop in the *1st* and the *2nd* trimer units seem to bend in similar manner, whereas that in the *3rd* trimer unit is clearly "streched up". This is in line with the observation that the peak in the V3 domain was lost or decresed in the RMSFs of the filtered trajectories. Additionally, this is in agreement with the COM distance calculations where the tip of the V3 loop of the *3rd* trimer unit clearly seemed to lie further away from the core.

Then, the range of the movement of the V1 loop in each system was visualized and is shown in the Figures 5.13(b) - 5.13(e). The pictures clearly show that the

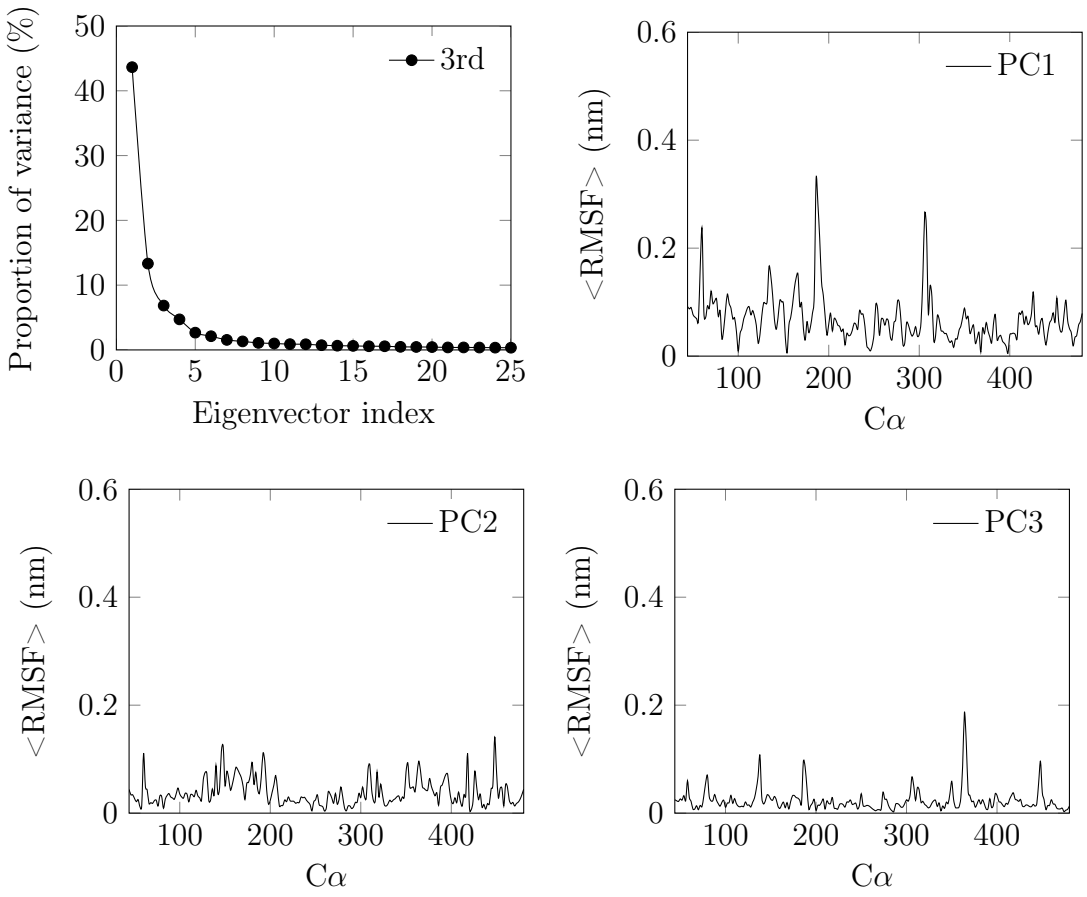
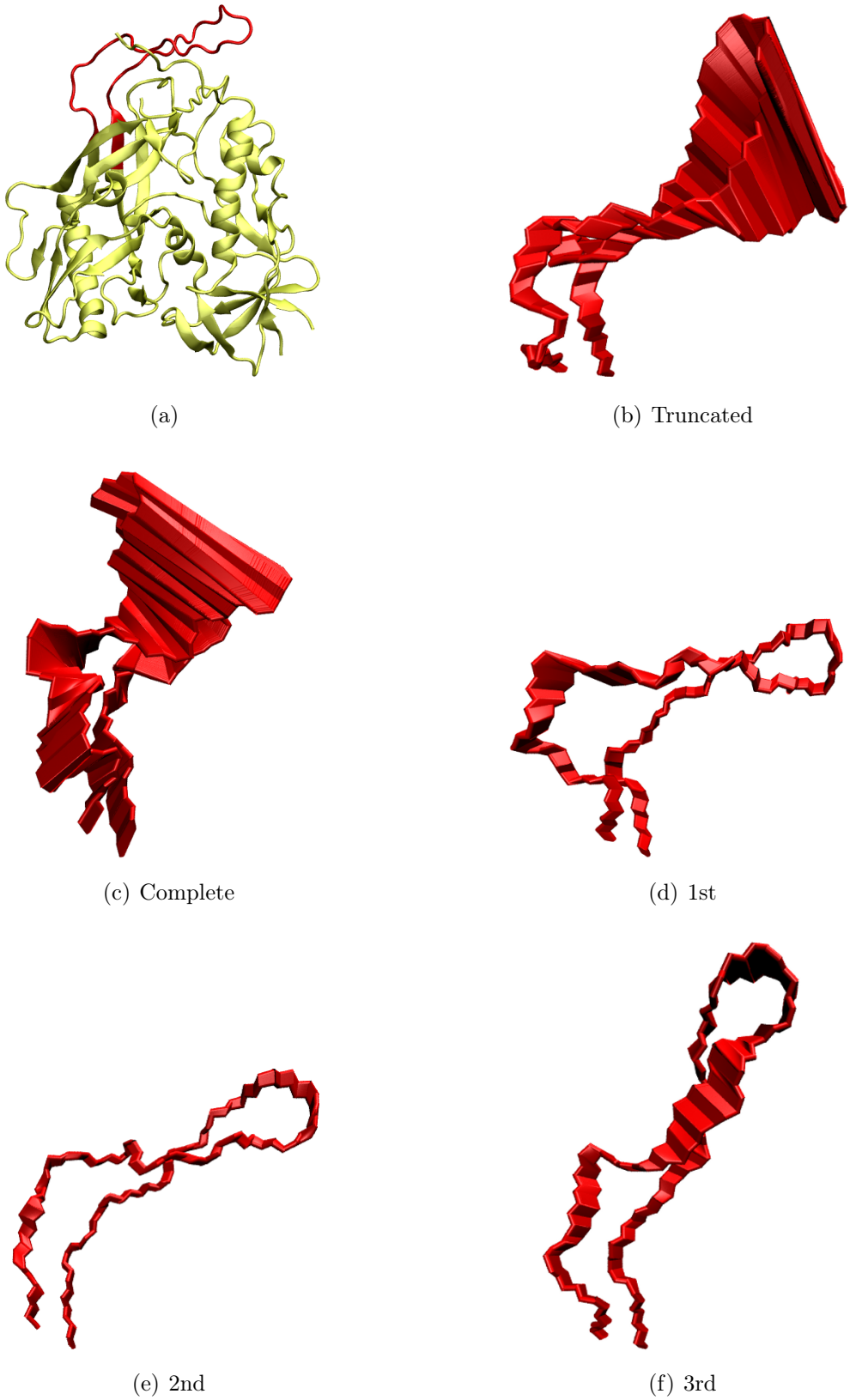


Figure 5.11 *The eigenvalues and the RMSFs from the principal components of the 3rd trimer unit.*

range is the greatest in the *complete* monomer and in the *2nd* trimer unit. This is again in agreement with the results from the principal component RMSFs of the *2nd* trimer unit shown in Figure 5.10, where there were clear peaks in the V1 domain. Additionally, it was found out before that the V1 loop of the *2nd* trimer unit was very flexible which is in line with the observation. Finally, the range of the movement of the V2 loop in each system was visualized and is shown in the Figures 5.14(b) - 5.14(e). The range of the movement in the *complete* monomer is clearly the widest which was expected as the RMSFs from the filtered trajectories showed clear peaks in this domain. In turn, the range of the movement in the V2 loop of the *2nd* trimer unit is very narrow. In comparison to this, the V2 loop of the *1st* and the *3rd* trimer units seem to move more along the first PC, however, not as much as in the monomer. Also in the RMSFs from the filtered trajectories there were peaks in these cases but they were significantly decreased in comparison to the *complete* monomer. Nonetheless, the pictures also show that the range of the movement of the V2 loop is rather different between the monomer and the trimer units. The V2 loop in the *complete* monomer might be, for example, more exposed.

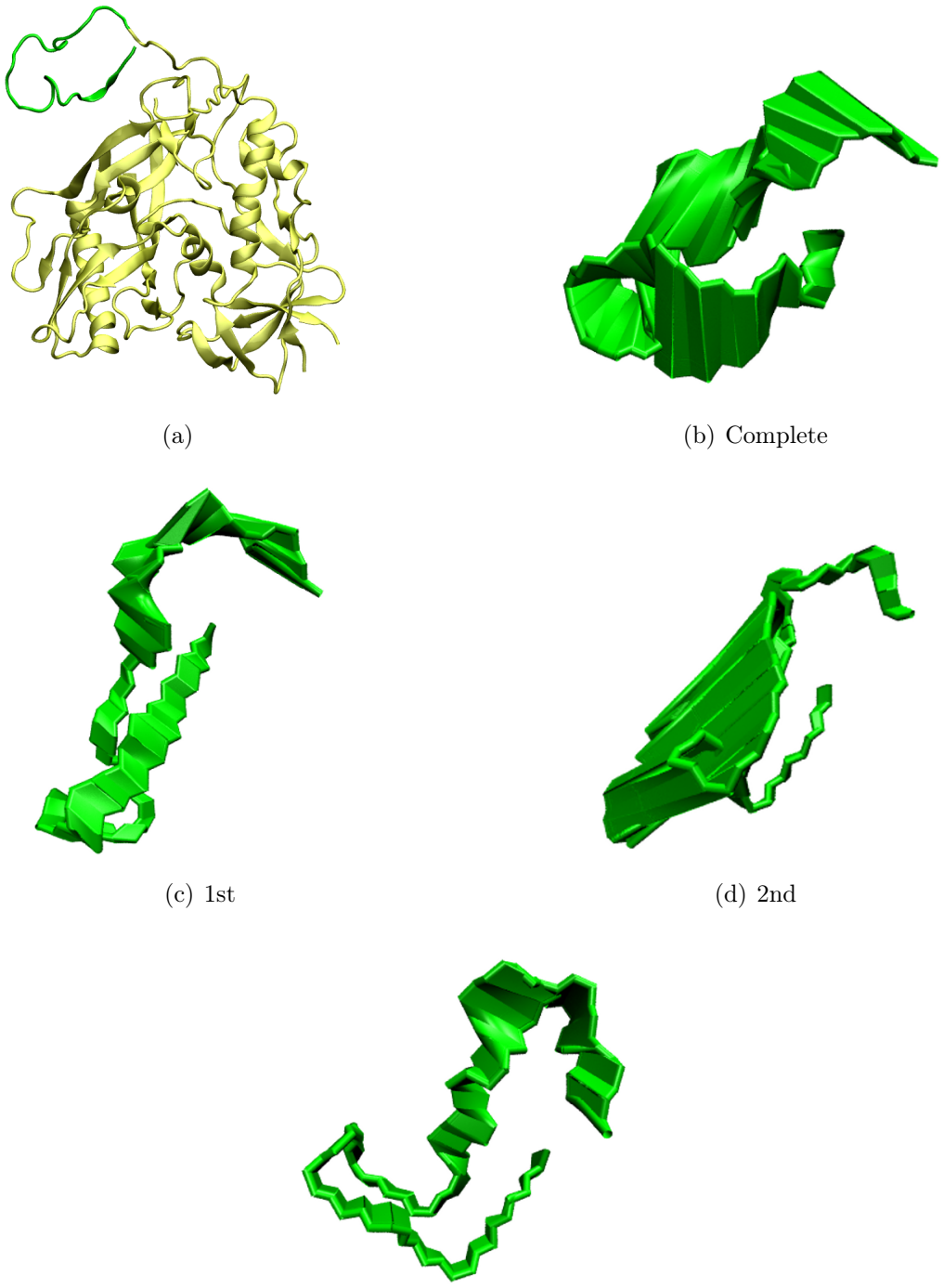


(a) (b) Truncated

(c) Complete (d) 1st

(e) 2nd (f) 3rd

Figure 5.12 *Range of movement of the V3 loops along the first principal component.*

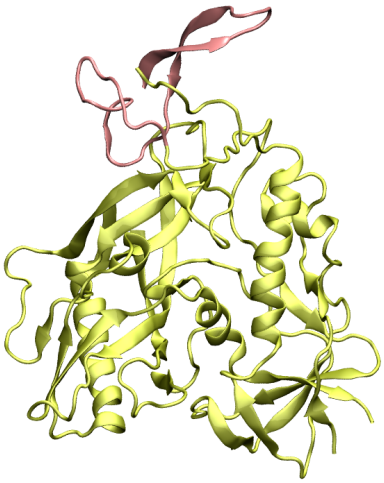


(a) (b) Complete

(c) 1st (d) 2nd

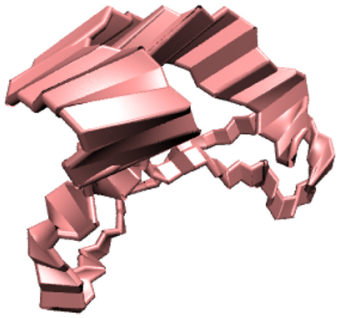
(e) 3rd

Figure 5.13 *Range of movement of the V1 loops along the first principal component.*



(a)

(b) Complete



(c) 1st

(d) 2nd

(e) 3rd

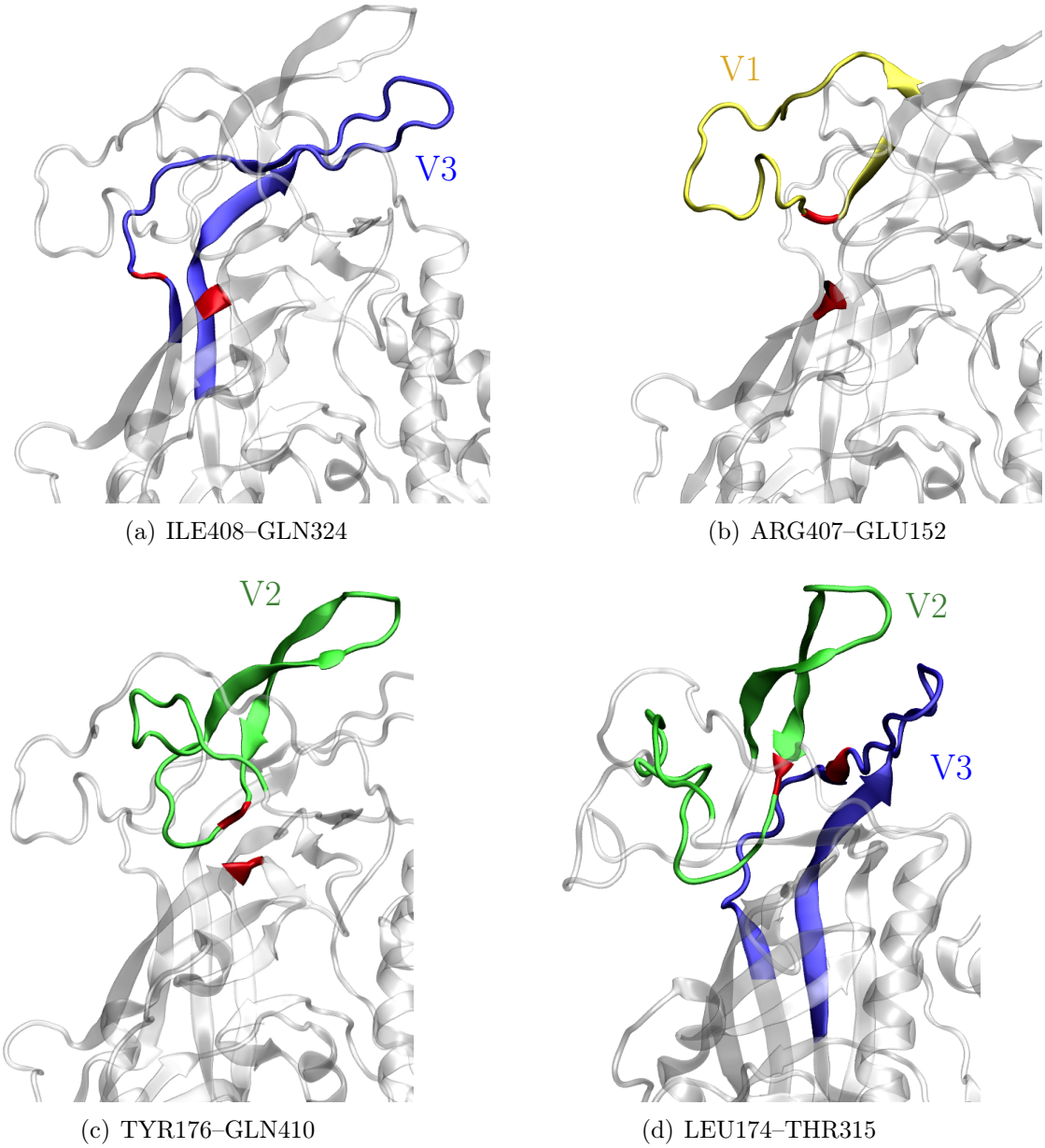
Figure 5.14 *Range of movement of the V2 loops along the first principal component.*

5.4 Inter-Loop Interactions Explain Reduced Mobility in Trimer

Next, noncovalent interactions within each system was studied. First, mean smallest distances between the residues were calculated in order to see which residues lie in close vicinity. For this, the GROMACS tool g_mdmat was used. The results were then visualized with color maps. The maps from each gp120 are shown in Figures C.1 - C.3. As might be expected, differences were especially seen between the V1/V2 and V3 domains. Then, hydrogen bonds and salt bridges between the core and these domains were calculated. A hydrogen bond is an electrostatic attraction between polar molecules. It occurs when a hydrogen atom binds to highly electronegative atoms, oxygen (O), nitrogen (N), and fluorine (F). Hydrogen bonds were calculated in VMD with the Hydrogen Bonds extension. The donor-acceptor distance was set at 0.325 nm and the cutoff angle at 35^o. A salt bridge, in turn, is a bond between oppositely charged residues that are sufficiently close to each other. Salt bridges occur between negatively and positively charged amino acids, that is, between Aspartic and Glutamatic acid (negative), and Arginine, Histidine, and Lysine (positive). Salt bridges were calculated in VMD with the Salt Bridges extension. A salt bridge was considered when the distance between an amide N and a carboxyl O was less than or equal to 0.45 nm. In all calculations the first 200 ns from the trajectories were excluded.

First, the average number of the hydrogen bonds with their standard deviations were calculated. The calculation was performed on between the core and each major variable loops seprately and then between the major variable loops. The results are shown in Table D.1. No remarkable differences were found between the systems. However, when the occurrence of certain bonds was studied in detail, some differences were found. The occurrence of certain hydrogen bonds and salt bridges between the core and V3 domain are shown in Tables D.2 and D.3. Those between the core and the V1/V2 domain in Table D.4. All bonds that existed at least 10 % of the time after the stabilization were considered. Then, similarities and remarkable differences between the systems were looked for.

Between the core and the V3 domain five bonds were found in all systems (marked in Tables D.2 and D.3). This indicates that rather many interactions remained between the V3 loop and the core regardless of the presence of the V1/V2 domain and the trimer conformation. In addition, there was one hydrogen bond that only existed in the *complete* monomer and in the *3rd* trimer unit. The disposition of this bond is shown in Figure 5.15(a). Between the core and the V1/V2 domain one bond was found in all systems (marked in Table D.4). Additionally, one bond was found in all



(a) ILE408–GLN324

(b) ARG407–GLU152

(c) TYR176–GLN410

(d) LEU174–THR315

Figure 5.15 *Hydrogen bonds that were only found (a) in the complete monomer and in the 3rd trimer unit (b) in the trimer units (c)-(d) in the 1st and the 2nd trimer units.*

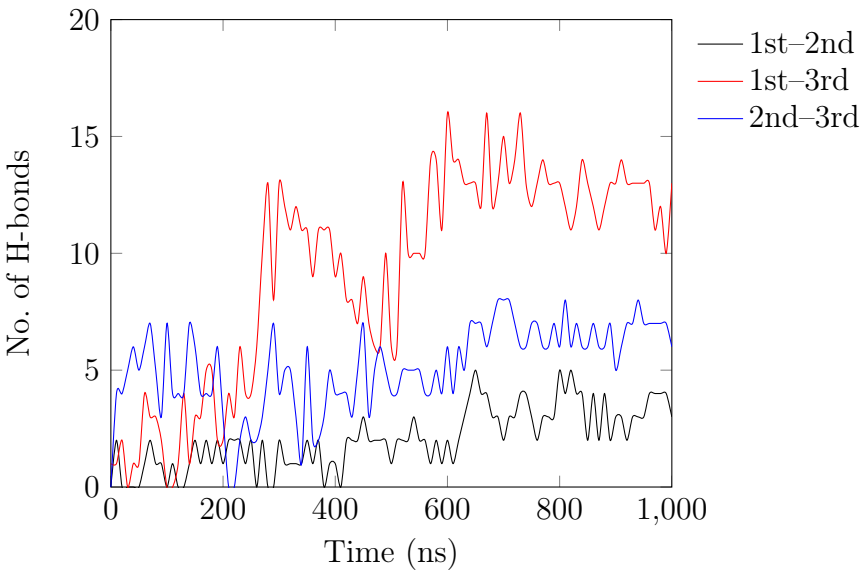


Figure 5.16 *Number of hydrogen bonds between the trimer units.*

trimer units but not in the monomer. The disposition of this bond is shown in Figure 5.15(b). One bond was also found that only existed in the *1st* and the *2nd* trimer units. The disposition of this bond is shown in Figure 5.15(c). Between the V1/V2 and V3 domains no common bonds between the systems were found. However, one hydrogen bond was only found in the *1st* and *2nd* trimer units. The disposition of this bond is shown in Figure 5.15(d). Nonetheless, even though several bonds were found that only existed in two or more systems, their effect on the conformation was not recognized here.

Then, hydrogen bonds between the trimer units were looked for. First, the total number of hydrogen bonds between the units were calculated. The time evolution of the hydrogen bonds is shown in Figure 5.16. It was found out that there were clearly the most hydrogen bonds between the *1st* and the *3rd* trimer unit. This indicates that the trimer units interacted asymmetrically in the simulation. Then, the occurrence of certain hydrogen bonds and salt bridges were studied between each variable loop and its adjacent gp120 units. The results are shown in Table D.6. It was found out that there were also the most residues taking part in the bonding between the *1st* and the *3rd* trimer units. One certain hydrogen bond appeared to exist between all trimer units (marked in Table 5.16), and in fact, it was found out to be the most permanent (84 % of the time) between the *1st* and the *3rd* trimer unit. Additionally, it was found out that the V2 and V3 loops were the most active domains in forming hydrogen bonds between the trimer units. Finally, three hydrogen bonds were found between a variable loop and its adjacent gp120 core. Two of them were found between the *1st* and the *3rd* trimer unit and the third

between the *2nd* and the *3rd* trimer units. All in all, as not big changes in the hydrogen bonding inside gp120 between monomer and trimer systems were found, the reduced motion of the major variable loops in the trimer was propably mostly due to inter-gp120 interactions.

5.5 V1/V2 Increases Structural Dynamics of V3 Loop in Monomer

Finally, the secondary structure of the major variable loops of the gp120 was calculated with the DSSP program [98]. GROMACS package provides an interface, do_dssp, that was used to run the program. First, the average number of residues with each secondary structure type and their standard deviations were calculated. It was found out that β-structures, especially β-sheets, were the most abundant structures present. The average number of residues in the loop domains possessing them are listed in Table E.1. Then, the secondary structure of the variable loops were calculated as a function of time. The secondary structures of the V3 loops are shown in the Figure E.1 and those of the V1/V2 loops in Figure E.2. It was found out that the secondary structure of the V3 loop of the *complete* monomer varied the most during the simulation. Accordingly, the secondary structure of the V1/V2 domain of the *complete* monomer varied the most during the simulation. This indicates that the loop structures of the *complete* monomer were the most dynamic. However, the secondary structure of the V3 loop of *truncated* monomer did not vary much. Hence, according to the simulations carried out here, the V1/V2 domain might increase the dynamics of V3 loop secondary structure in the monomer conformation. There were also differences in the secondary structure of the loops in the trimer: The V3 loop of the *1st* gp120 subunit had the least β-sheets. The V1/V2 domain of the *2nd* gp120 subunit of the trimer had the least β-sheets. However, there were not remarkable dynamic changes. Instead, the variable loops of the trimer mostly maintained constant secondary structures during the simulation.

6. CONCLUSIONS

In this Thesis, the conformational dynamics of functionally important loops of HIV-1 gp120 was studied with the aid of atomistic molecular dynamics (MD) simulations. Gp120 is the main target for vaccines against HIV infection. In experiments only the monomeric gp120 is often considered, even though the native state of the protein is a trimer. Additionally, the functionally important variable loops (V1, V2, and V3) are often unresolved or excluded in experimental studies. What is more, experiments have only been able to give static pictures of the protein even though it is known that the protein functions often arise from its dynamics. There are some atomistic molecular dynamics studies that have managed to enlighten gp120 dynamics [5, 6, 7, 90, 91, 92]. However, all these MD studies have only considered the monomeric gp120. In addition, the V1 and V2 loops have not been included in the gp120 structure.

In this MD study, these shortcomings have been overcome. All variable loops of gp120, including the previously missing V1 and V2 loops, have been included in the systems. Additionally, the trimeric gp120 has been considered. The preparation of the systems was carried out by combining known crystal structures of gp120 and by following the sequence of a native HIV-1 isolate YU-2. In all, three systems were built: First, a monomeric gp120 core with only *three* variable loops, V3 to V5. Then, a monomeric gp120 core with *all* variable loops, V1 to V5. Finally, a trimer of three gp120 subunits with *all* variable loops, V1 to V5, and three gp41 subunits that are needed for the trimer stability. With these systems it was possible to study whether the V1/V2 domain inclusion affects gp120 dynamics and if the dynamics change in a trimer context. The systems were all dissolved in water with a physiolocigal salt concentration and simulated for 1 μs.

It is known that the gp120 core is a rather stable structure [4, 17]. It does not deviate much even if gp120 binds to different receptors and antibodies. It is also known that in the monomeric gp120 the V3 loop is extremely flexible and plays key roles in the functions of gp120, for example, in determining the co-receptor speficificity [90]. In this study, similar observations of the V3 loop flexibility were made. The V3 loop was found out to be very flexible in both gp120 monomers despite the

presence of the V1/V2 domain. However, in the trimer the V3 loop was not found to be flexible anymore. Similarly, the V1/V2 domain was discovered to be very flexible in the monomer but generally not in the trimer. In addition, all major loops showed dynamic mobility during the monomer simulations, whereas in the trimer simulation the loop mobility was significantly reduced. Conformational distribution of the structure during the simulations also verified the gp120 monomer dynamics in comparison with the trimeric gp120. Then, the principal component analysis showed that there was prominent concerted motion in the monomer between the V1/V2 and V3 domains. However, the concerted motion was mostly lost in the trimer. When the range of the movement of the major variable loops along the first principal component, i.e. the direction of greatest variance, was visualised, it was found out that especially the V3 loop movement was significantly narrower in the trimer than in the monomer. Similarly, the range of the movement of the V1 and V2 loops along the first principal component was also generally narrower but was more clearly seen in some trimer units than in the others.

To investigate the causes of the changes in gp120 variable loop dynamics, intra- and inter-gp120 hydrogen bonds and salt bridges were looked for. Additionally, the secondary structure of the loops was studied. Some changes were found inside gp120 in hydrogen bonding. However, the most probable causes for the reduced flexibility, mobility, and lost concerted motion in the trimer were the inter-gp120 interactions. It was found out that especially the variable loops V2 and V3 actively formed hydrogen bonds with their adjacent gp120 variable loops in the trimer. Previous structural studies also suggest that the gp120 subunits are held together, at least in part, by association of the V1, V2, and V3 regions at the apex of the trimer [99, 100, 55]. Additionally, it was found out that the interaction between the trimer units was asymmetrical. Two of the three gp120 subunits interacted significantly more via hydrogen bonding. This probably explains small differences between the gp120 subunits that were seen throughout the analysis.

Nonetheless, the most crucial findings in this study were the differences in the variable loop dynamics of the monomeric and trimeric gp120. The V3 loop has been shown to be well exposed in monomeric gp120 in the presence of the CD4 receptor [30, 101, 102]. In addition, atomistic MD simulation studies suggest high flexibility in the V3 domain and it has been proposed that the plasticity of the loop is crucial for the CD4 receptor binding [90]. However, it has been unclear whether the V3 loop is similarly extended on the trimers of the viral spike.

According to the simulations ran for this Thesis, the V3 loop is neither extended nor flexible in the trimer. Instead, numerous hydrogen bond interactions between

the variable loops of gp120 subunits of the trimer hold the loop at rest on the trimer spike. In these interactions, the variable loops V2 and V3 play key roles in stabilizing the trimer apex. Similarly and due to the same interactions, the V1/V2 domain has partly lost its mobility in the trimer spike. This suggests that the major variable loops, especially the V3 loop, do not necessarily need to be exposed before CD4 binding and that Env has to undergo significant conformational changes in order to have the loop region accessible for receptor binding. These changes are achieved either by inherent conformational dynamics or are induced by the receptor.

For the sake of the reliability of the results, MD simulations carried out for this Thesis need to be repeated. Additionally, one crucial factor should be taken into consideration. The viral spike of HIV-1 is known to be highly glycosylated. Half of the mass of gp120 comes from its glycan shield [4]. Until now, experimental methods have been rather powerless in studying these flexible and varying carbohydrate structures that significantly mask the gp120 surface and hinder the binding of antibodies. Most importantly, the effects of the glycans on the protein dynamics are poorly understood. Instead, MD has great potential in this regard as various permanent glycan binding sites on gp120 are known. In fact, some progress in studying the influence of the glycans in the V3 loop dynamics has already been done [5]. What is more, MD simulation systems representing native HIV-1 isolates, other than YU-2 used here, would aid in development of a general model for the viral entry mechanism and most importantly in developing vaccines against the virus.

BIBLIOGRAPHY

[1] F. Barré-Sinoussi, J.-C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vézinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. Isolation of a T-lymphotropic Retrovirus from a Patient at Risk for Acquired Immune Deficiency Syndrome (AIDS). *Science*, 220:868–871, 1983.

[2] P. J. Kanki, K. U. Travers, R. G. Marlink, M. E. Essex, S. MBoup, A. Gueye-NDiaye, T. Siby, I. Thior, J.-L. Sankalé, C.-C. Hsieh, M. Hernandez-Avila, and I. NDoye. Slower Heterosexual Spread of HIV-2 than HIV-1. *The Lancet*, 343:943–946, 1994.

[3] A. Bartesaghi, A. Merk, M. J. Borgnia, J. L. S. Milne, and S. Subramaniam. Prefusion Structure of Trimeric HIV-1 Envelope Glycoprotein Determined by Cryo-electron Microscopy. *Nature Structural & Molecular Biology*, 20:1352–1357, 2013.

[4] R. Pantophlet and D. R. Burton. Gp120: Target for Neutralizing HIV-1 Antibodies. *Annual Review of Immunology*, 24:739–769, 2006.

[5] N. T. Wood, E. Fadda, R. Davis, O. C. Grant, J. C. Martin, R. J. Woods, and S. A. Travers. The Influence of N-Linked Glycans on the Molecular Dynamics of the HIV-1 Gp120 V3 Loop. *PLoS One*, 8:e80301, 2013.

[6] P. Sang, L.-Q. Yang, X.-L. Ji, Y.-X. Fu, and S.-Q. Liu. Insight Derived from Molecular Dynamics Simulations into Molecular Motions, Thermodynamics and Kinetics of HIV-1 Gp120. *PLoS One*, 9:e104714, 2014.

[7] M. Yokoyama, S. Naganawa, K. Yoshimura, S. Matsushita, and H. Sato. Structural Dynamics of HIV-1 Envelope Gp120 Outer Domain with V3 Loop. *PLoS One*, 7:e37530, 2012.

[8] M. D. Daniel, N. L. Letvin, N. W. King, M. Kannagi, P. K. Sehgal, R. D. Hunt, P. J. Kanki, M. Essex, and R. C. Desrosiers. Isolation of T-cell Tropic HTLV-III-Like Retrovirus from Macaques. *Science*, 228:1201–1204, 1985.

[9] R. C. Gallo, S. Z. Salahuddin, M. Popovic, G. M. Shearer, M. Kaplan, B. F. Haynes, T. J. Palker, R. Redfield, J. Oleske, B. Safai, et al. Frequent Detection and Isolation of Cytopathic Retroviruses (HTLV-III) from Patients with AIDS and at Risk for AIDS. *Science*, 224:500–503, 1984.

[10] N. L. Letvin, M. D. Daniel, P. K. Sehgal, R. C. Desrosiers, R. D. Hunt, L. M. Waldron, J. J. MacKey, D. K. Schmidt, L. V. Chalifoux, and N. W. King. Induction of AIDS-like Disease in Macaque Monkeys with T-cell Tropic Retrovirus STLV-III. *Science*, 230:71–73, 1985.

[11] R. Wyatt and J. Sodroski. The HIV-1 Envelope Glycoproteins: Fusogens, Antigens, and Immunogens. *Science*, 280:1884–1888, 1998.

[12] D. L. Robertson, J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, M. L. Kalish, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky, and B. Korber. HIV-1 Nomenclature Proposal. *Science*, 288:55–55, 2000.

[13] J. Hemelaar, E. Gouws, P. D. Ghys, and S. Osmanov. Global and Regional Distribution of HIV-1 Genetic Subtypes and Recombinants in 2004. *Aids*, 20:W13–W23, 2006.

[14] E. Tschachler. The Dermatologist and the HIV/AIDS Pandemic. *Clinics in Dermatology*, 32:286–289, 2014.

[15] S. Zolla-Pazner. A Critical Question for HIV Vaccine Development: Which Antibodies to Induce? *Science*, 345:167–168, 2014.

[16] K. Fenstermacher. The mature HIV virion. figshare. [online], http://dx.doi.org/10.6084/m9.figshare.862069, 2013.

[17] A. Merk and S. Subramaniam. HIV-1 Glycoprotein Structure. *Current Opinion in Structural Biology*, 23:268–276, 2013.

[18] W. Weissenhorn, A. Dessen, L. J. Calder, S. C. Harrison, J. J. Skehel, and D. C. Wiley. Structural Basis for Membrane Fusion by Enveloped Viruses. *Molecular Membrane Biology*, 16:3–9, 1999.

[19] P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson. Structure of an HIV Gp120 Envelope Glycoprotein in Complex with the CD4 Receptor and a Neutralizing Human Antibody. *Nature*, 393:648–659, 1998.

[20] M. Kowalski, J. Potz, L. Basiripour, T. Dorfman, W. C. Goh, E. Terwilliger, A. Dayton, C. Rosen, W. Haseltine, and J. Sodroski. Functional Regions of the Envelope Glycoprotein of Human Immunodeficiency Virus Type 1. *Science*, 237:1351–1355, 1987.

[21] S. A. Gallo, C. M. Finnegan, M. Viard, Y. Raviv, A. Dimitrov, S. S. Rawat, A. Puri, S. Durell, and R. Blumenthal. The HIV Env-Mediated Fusion Reaction. *Biochimica et Biophysica Acta Biomembranes*, 1614:36–50, 2003.

[22] M. Lu, S. C. Blacklow, and P. S. Kim. A Trimeric Structural Domain of the HIV-1 Transmembrane Glycoprotein. *Nature Structural & Molecular Biology*, 2:1075–1082, 1995.

[23] D. C. Chan, D. Fass, J. M. Berger, and P. S. Kim. Core Structure of Gp41 from the HIV Envelope Glycoprotein. *Cell*, 89:263–273, 1997.

[24] W. Weissenhorn, A. Dessen, S. C. Harrison, J. J. Skehel, and D. C. Wiley. Atomic Structure of the Ectodomain from HIV-1 Gp41. *Nature*, 387:426–430, 1997.

[25] K. Tan, J.-H. Liu, J.-H. Wang, S. Shen, and M. Lu. Atomic Structure of a Thermostable Subdomain of HIV-1 Gp41. *Proceedings of the National Academy of Sciences USA*, 94:12303–12308, 1997.

[26] B. R. Starcich, B. H. Hahn, G. M Shaw, P. D. McNeely, S. Modrow, H. Wolf, E. S. Parks, W. P. Parks, S. F. Josephs, R. C. Gallo, and F. Wong-Staal. Identification and Characterization of Conserved and Variable Regions in the Envelope Gene of HTLV-III/LAV, the Retrovirus of AIDS. *Cell*, 45:637–648, 1986.

[27] J. P. Moore, Q. J. Sattentau, R. Wyatt, and J. Sodroski. Probing the Structure of the Human Immunodeficiency Virus Surface Glycoprotein Gp120 with a Panel of Monoclonal Antibodies. *Journal of Virology*, 68:469–484, 1994.

[28] R. Wyatt, N. Sullivan, M. Thali, H. Repke, D. Ho, J. Robinson, M. Posner, and J. Sodroski. Functional and Immunologic Characterization of Human Immunodeficiency Virus Type 1 Envelope Glycoproteins Containing Deletions of the Major Variable Regions. *Journal of Virology*, 67:4557–4565, 1993.

[29] S. R. Pollard, M. D. Rosa, J. J. Rosa, and D. C. Wiley. Truncated Variants of Gp120 Bind CD4 with High Affinity and Suggest a Minimum CD4 Binding Region. *The EMBO Journal*, 11:585, 1992.

[30] C.-C. Huang, M. Tang, M.-Y. Zhang, S. Majeed, E. Montabana, R. L. Stanfield, D. S. Dimitrov, B. Korber, J. Sodroski, I. A. Wilson, R. Wyatt, and P. D. Kwong. Structure of a V3-containing HIV-1 Gp120 Core. *Science*, 310:1025–1028, 2005.

[31] R. Pejchal, K. J. Doores, L. M. Walker, R. Khayat, P.-S. Huang, S.-K. Wang, R. L. Stanfield, J.-P. Julien, A. Ramos, M. Crispin, R. Depetris, U. Katpally, A. Marozsan, A. Cupo, S. Maloveste, Y. Liu, R. McBride, Y. Ito, R. W. Sanders, C. Ogohara, J. C. Paulson, T. Feizi, C. N. Scanlan, C. H. Wong, J. P. Moore, W. C. Olson, A. B. Ward, P. Poignard, W. R. Schief, D. R. Burton, and I. A. Wilson. A Potent and Broad Neutralizing Antibody Recognizes and Penetrates the HIV Glycan Shield. *Science*, 334:1097–1103, 2011.

[32] M. Pancera, S. Majeed, Y.-E. A. Ban, L. Chen, G.-C. Huang, L. Kong, Y. Do Kwon, J. Stuckey, T. Zhou, J. E Robinson, W. R. Schief, R. Sodroski, J. Wyatt, and P. D. Kwong. Structure of HIV-1 Gp120 with Gp41-Interactive Region Reveals Layered Envelope Architecture and Basis of Conformational Mobility. *Proceedings of the National Academy of Sciences USA*, 107:1166–1171, 2010.

[33] J.-P. Julien, A. Cupo, D. Sok, R. L. Stanfield, D. Lyumkis, M. C. Deller, P.-J. Klasse, D. R. Burton, R. W. Sanders, J. P. Moore, and Wilson I. A. Ward, A. B. Crystal Structure of a Soluble Ceaved HIV-1 Envelope Trimer. *Science*, 342:1477–1483, 2013.

[34] D. N. Marti, S. Bjelić, M. Lu, H. R. Bosshard, and I. Jelesarov. Fast Folding of the HIV-1 and SIV Gp41 Six-Helix Bundles. *Journal of Molecular Biology*, 336:1–8, 2004.

[35] P. D. Kwong, M. L. Doyle, D. J. Casper, C. Cicala, S. A. Leavitt, S. Majeed, T. D. Steenbeke, M. Venturi, I. Chaiken, M. Fung, H. Katinger, P. W. Parren, J. Robinson, R. D. Van, L. Wang, D. R. Burton, E. Freire, R Wyatt, J. Sodroski, W. A. Hendrickson, and J. Arthos. HIV-1 Evades Antibody-Mediated Neutralization through Conformational Masking of Receptor-Binding Sites. *Nature*, 420:678–682, 2002.

[36] R. Wyatt, P. D. Kwong, E. Desjardins, R. W. Sweet, J. Robinson, W. A. Hendrickson, and J. G. Sodroski. The Antigenic Structure of the HIV Gp120 Envelope Glycoprotein. *Nature*, 393:705–711, 1998.

[37] C. D. Rizzuto, R. Wyatt, N. Hernández-Ramos, Y. Sun, P. D. Kwong, W. A. Hendrickson, and J. Sodroski. A Conserved HIV Gp120 Glycoprotein Structure Involved in Chemokine Receptor Binding. *Science*, 280:1949–1953, 1998.

[38] T. Cardozo, T. Kimura, S. Philpott, B. Weiser, H. Burger, and S. Zolla-Pazner. Structural Basis for Coreceptor Selectivity by the HIV Type 1 V3 Loop. *AIDS Research and Human Retroviruses*, 23:415–426, 2007.

[39] C. R. Bertozzi and L. L. Kiessling. Chemical Glycobiology. *Science*, 291:2357–2364, 2001.

[40] A. Helenius and M. Aebi. Intracellular Functions of N-Linked Glycans. *Science*, 291:2364–2369, 2001.

[41] C. K. Leonard, M. W. Spellman, L. Riddle, R. J. Harris, J. N. Thomas, and T. J. Gregory. Assignment of Intrachain Disulfide Bonds and Characterization of Potential Glycosylation Sites of the Type 1 Recombinant Human Immunodeficiency Virus Envelope Glycoprotein (Gp120) Expressed in Chinese Hamster Ovary Cells. *Journal of Biological Chemistry*, 265:10373–10382, 1990.

[42] X Zhu, C. Borchers, R. J. Bienstock, and K. B. Tomer. Mass Spectrometric Characterization of the Glycosylation Pattern of HIV-Gp120 Expressed in CHO Cells. *Biochemistry*, 39:11194–11204, 2000.

[43] A. Land and I. Braakman. Folding of the Human Immunodeficiency Virus Type 1 Envelope Glycoprotein in the Endoplasmic Reticulum. *Biochimie*, 83:783–790, 2001.

[44] Y. Li, L. Luo, N. Rasool, and C. Y. Kang. Glycosylation Is Necessary for the Correct Folding of Human Immunodeficiency Virus Gp120 in CD4 Binding. *Journal of Virology*, 67:584–588, 1993.

[45] D. Wilhelm, H. N Behnken, and B. Meyer. Glycosylation Assists Binding of HIV Protein Gp120 to Human CD4 Receptor. *ChemBioChem*, 13:524–527, 2012.

[46] G. Pollakis, S. Kang, A. Kliphuis, M. I. M. Chalaby, J. Goudsmit, and W. A. Paxton. N-Linked Glycosylation of the HIV Type-1 Gp120 Envelope Glycoprotein as a Major Determinant of CCR5 and CXCR4 Coreceptor Utilization. *Journal of Biological Chemistry*, 276:13433–13441, 2001.

[47] X. Wei, J. M. Decker, S. Wang, H. Hui, J. C. Kappes, X. Wu, J. F Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A. Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw. Antibody Neutralization and Escape by HIV-1. *Nature*, 422:307–312, 2003.

[48] E. J. Toone. Structure and Energetics of Protein-Carbohydrate Complexes. *Current Opinion in Structural Biology*, 4:719–728, 1994.

[49] L. M. Walker and D. R. Burton. Rational Antibody-based HIV-1 Vaccine Design: Current Approaches and Future Directions. *Current Opinion in Immunology*, 22:358–366, 2010.

[50] M. E. Curlin, R. Zioni, S. E. Hawes, Y. Liu, W. Deng, G. S. Gottlieb, T. Zhu, and J. I. Mullins. HIV-1 Envelope Subregion Length Variation During Disease Progression. *PLoS Pathogens*, 6:e1001228, 2010.

[51] R. F Speck, K. Wehrly, E. J. Platt, R. E. Atchison, I. F. Charo, D. Kabat, B. Chesebro, and M. A. Goldsmith. Selective Employment of Chemokine Receptors as Human Immunodeficiency Virus Type 1 Coreceptors Determined by Individual Amino Acids within the Envelope V3 Loop. *Journal of Virology*, 71:7136–7139, 1997.

[52] M. J. van Gils, E. M. Bunnik, B. D. Boeser-Nunnink, J. A. Burger, M. Terlouw-Klein, N Verwer, and H. Schuitemaker. Longer V1V2 Region with Increased Number of Potential N- Linked Glycosylation Sites in the HIV-1 Envelope Glycoprotein Protects Against HIV-Specific Neutralizing Antibodies. *Journal of Virology*, 85:6986–6995, 2011.

[53] J. S. McLellan, M. Pancera, C. Carrico, J. Gorman, J.-P. Julien, R. Khayat, R. Louder, R. Pejchal, M. Sastry, K. Dai, S. OĎell, N. Patel, S. Shahzad-ul Hussan, Y. Yang, B. Zhang, T. Zhou, J. Zhu, J. C. Boyington, G. Y. Chuang, D. Diwanji, I. Georgiev, Y. D. Kwon, D. Lee, M. K. Louder, S. Moquin, S. D. Schmidt, Z. Y. Yang, M. Bonsignori, J. A. Crump, S. H. Kapiga, N. E. Sam, B. F. Haynes, D. R. Burton, W. C. Koff, L. M. Walker, S. Phogat, R. Wyatt, J. Orwenyo, L. X. Wang, J. Arthos, C. A. Bewley, J. R. Mascola, G. J. Nabel, W. R. Schief, A. B. Ward, I. A. Wilson, and P. D. Kwong. Structure of HIV-1 Gp120 V1/V2 Domain with Broadly Neutralizing Antibody PG9. *Nature*, 480:336–343, 2011.

[54] Y. Do Kwon, A. Finzi, X. Wu, C. Dogo-Isonagie, L. K. Lee, L. R. Moore, S. D. Schmidt, J. Stuckey, Y. Yang, T. Zhou, J. Zhu, D. A. Vicic, A. K. Debnath, L. Shapiro, C. A. Bewley, Sodroski J. G. Mascola, J. R., and P. D. Kwong. Unliganded HIV-1 Gp120 Core Structures Assume the CD4-Bound Conformation with Regulation by Quaternary Interactions and Variable Loops. *Proceedings of the National Academy of Sciences USA*, 109:5663–5668, 2012.

[55] J. Liu, A. Bartesaghi, M. J. Borgnia, G. Sapiro, and S. Subramaniam. Molecular Architecture of Native HIV-1 Gp120 Trimers. *Nature*, 455:109–113, 2008.

[56] E. E. H. Tran, M. J. Borgnia, O. Kuybeda, D. M. Schauder, A. Bartesaghi, G. A. Frank, G. Sapiro, J. L. S. Milne, and S. Subramaniam. Structural Mechanism of Trimeric HIV-1 Envelope Glycoprotein Activation. *PLoS Pathogens*, 8:e1002797, 2012.

[57] M. Pancera, T. Zhou, A. Druz, I. S. Georgiev, C. Soto, J. Gorman, J. Huang, P. Acharya, G.-Yu. Chuang, G. Ofek, B. E. Stewart-Jones, J. Stuckey, R. T. Bailer, M. G. Joyce, M. K. Louder, N. Tumba, Y. Yang, B. Zhang, M. S. Cohen, B. F. Haynes, J. R. Mascola, L. Morris, J. B. Munro, S. C. Blanchard, W. Mothes, M Connors, and P. D. Kwong. Structure and Immune Recognition of Trimeric Pre-Fusion HIV-1 Env. *Nature*, 514:455–461, 2014.

[58] J. J. Skehel and D. C. Wiley. Coiled Coils in Both Intracellular Vesicle and Viral Membrane Fusion. *Cell*, 95:871–874, 1998.

[59] C. Grewe, A. Beck, and H. R. Gelderblom. HIV: Early Virus-Cell Interactions. *Journal of Acquired Immune Deficiency Syndromes*, 3:965–974, 1990.

[60] D. C. Chan and P. S. Kim. HIV Entry and Its Inhibition. *Cell*, 93:681–684, 1998.

[61] B. D. Welch, A. P. VanDemark, A. Heroux, C. P. Hill, and M. S.S Kay. Potent D-Peptide Inhibitors of HIV-1 Entry. *Proceedings of the National Academy of Sciences USA*, 104:16828–16833, 2007.

[62] D. Klatzmann, E. Champagne, S. Chamaret, J. Gruest, D. Guetard, T. Hercend, J.-C. Gluckman, and L. Montagnier. T-lymphocyte T4 Molecule Behaves as the Receptor for Human Retrovirus LAV. *Nature*, 312, 1984.

[63] J. S. McDougal, M. S. Kennedy, J. M. Sligh, S. P. Cort, A. Mawle, and J. K. Nicholson. Binding of HTLV-III/LAV to T4+ T Cells by a Complex of the 110K Viral Protein and the T4 Molecule. *Science*, 231:382–385, 1986.

[64] G. Alkhatib, C. Combadiere, C. C. Broder, Y. Feng, P. E Kennedy, P. M. Murphy, and E. A. Berger. CC CKR5: a RANTES, MIP-1α, MIP-1β Receptor as a Fusion Cofactor for Macrophage-Tropic HIV-1. *Science*, 272:1955–1958, 1996.

[65] H. Choe, M. Farzan, Y. Sun, N. Sullivan, B. Rollins, P. D. Ponath, L. Wu, C. R. Mackay, G. LaRosa, W. Newman, N. Gerard, C. Gerard, and J. Sodroski. The β-Chemokine Receptors CCR3 and CCR5 Facilitate Infection by Primary HIV-1 Isolates. *Cell*, 85:1135–1148, 1996.

[66] Y. Feng, C. C. Broder, P. E. Kennedy, and E. A. Berger. HIV-1 Entry Cofactor: Functional cDNA Cloning of a Seven-Transmembrane, G Protein-Coupled Receptor. *Science*, 272:872–877, 1996.

[67] Q. J. Sattentau and J. P. Moore. Conformational Changes Induced in the Human Immunodeficiency Virus Envelope Glycoprotein by Soluble CD4 Binding. *Journal of Experimental Medicine*, 174:407–415, 1991.

[68] M. Thali, J. P. Moore, C. Furman, M. Charles, D. D. Ho, J. Robinson, and J. Sodroski. Characterization of Conserved Human Immunodeficiency Virus Type 1 Gp120 Neutralization Epitopes Exposed upon Gp120-CD4 Binding. *Journal of Virology*, 67:3978–3988, 1993.

[69] H. Deng, R. Liu, W. Ellmeier, S. Choe, D. Unutmaz, M. Burkhart, P D. Marzio, S. Marmon, R. E Sutton, C. M. Hill, C. B. Davis, S. C. Peiper, T. J. Schall, Littman D. R., and N. R. Landau. Identification of a Major Co-Receptor for Primary Isolates of HIV-1. *Nature*, 381, 1996.

[70] S. A. Gallo, A. Puri, and R. Blumenthal. HIV-1 Gp41 Six-Helix Bundle Formation Occurs Rapidly after the Engagement of Gp120 by CXCR4 in the HIV-1 Env-Mediated Fusion Process. *Biochemistry*, 40:12231–12236, 2001.

[71] T. Schlick. *Molecular Modeling and Simulation: An Interdisciplinary Guide*, volume 21. Springer, New York, 2010.

[72] M. Karplus and J. A. McCammon. Molecular Dynamics Simulations of Biomolecules. *Nature Structural & Molecular Biology*, 9:646–652, 2002.

[73] T. Hansson, C. Oostenbrink, and W. van Gunsteren. Molecular Dynamics Simulations. *Current Opinion in Structural Biology*, 12:190–196, 2002.

[74] I. Vattulainen and T: Róg. Lipid Simulations: A Perspective on Lipids in Action. *Cold Spring Harbor Perspectives in Biology*, 3:a004655, 2011.

[75] D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, T. M. Sijbers, K. A. Feenstra, R. Drunen, and H. J. C. Berendsen. Gromacs User Manual Version 4.5.4. [online], http://www.gromacs.org, 2010.

[76] A. R. Leach. *Molecular Modelling: Principles and Applications*. Pearson Education, Harlow, 2nd edition, 2001.

[77] D. Frenkel and B. Smit. *Understanding Molecular Simulation: from Algorithms to Applications*. Academic Press, San Diego, 2nd edition, 2001.

[78] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw. Systematic Validation of Protein Force Fields Against Experimental Data. *PLoS One*, 7:e32131, 2012.

[79] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular Dynamics with Coupling to an External Bath. *Journal of Chemical Physics*, 81:3684–3690, 1984.

[80] S. Nosé. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Molecular Physics*, 52:255–268, 1984.

[81] W. G. Hoover. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Physical Review A*, 31:1695, 1985.

[82] G. Bussi, D. Donadio, and M. Parrinello. Canonical Sampling through Velocity Rescaling. *Journal of Chemical Physics*, 126:014101, 2007.

[83] M. Parrinello and A. Rahman. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *Journal of Applied Physics*, 52:7182–7190, 1981.

[84] G. J. Martyna, M. E. Tuckerman, D. J. Tobias, and M. L. Klein. Explicit Reversible Integrators for Extended Systems Dynamics. *Molecular Physics*, 87:1117–1157, 1996.

[85] G. Salvato-Vallverdu. Periodic Boundary Conditions. [online], http://www.texample.net/media/tikz/examples/PDF/periodic-boundaries-conditions.pdf, 2009.

[86] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A Smooth Particle Mesh Ewald Method. *Journal of Chemical Physics*, 103:8577–8593, 1995.

[87] L. Verlet. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, 159:98, 1967.

[88] R. W. Hockney. Potential Calculations and Some Applications. *Methods in Computational Physics*, 9:135–211, 1970.

[89] M. Paloncýová, G. Fabre, R. H. DeVane, P. Trouillas, K. Berka, and M. Otyepka. Benchmarking of Force Fields for Molecule–Membrane Interactions. *Journal of Chemical Theory and Computation*, 10:4143–4151, 2014.

[90] C. Balasubramanian, G. Chillemi, I. Abbate, M. R. Capobianchi, G. Rozera, and A. Desideri. Importance of V3 loop flexibility and net charge in the context of co-receptor recognition. A molecular dynamics study on HIV Gp120. *Journal of Biomolecular Structure and Dynamics*, 29:879–891, 2012.

[91] Y. Pan, B. Ma, and R. Nussinov. CD4 Binding Partially Locks the Bridging Sheet in Gp120 But Leaves the $\beta 2/3$ Strands Flexible. *Journal of Molecular Biology*, 350:514–527, 2005.

[92] L.-T. Da, J.-M. Quan, and Y.-D. Wu. Understanding of the Bridging Sheet Rormation of HIV-1 Glycoprotein Gp120. *Journal of Physical Chemistry B*, 113:14536–14543, 2009.

[93] P. D. Kwong, R. Wyatt, S. Majeed, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson. Structures of HIV-1 Gp120 Envelope Glycoproteins from Laboratory-Adapted and Primary Isolates. *Structure*, 8:1329–1339, 2000.

[94] P. Acharya, T. S. Luongo, M. K. Louder, K. McKee, Y. Yang, Y. Do Kwon, J. R. Mascola, P. Kessler, L. Martin, and P. D. Kwong. Structural Basis for Highly Effective HIV-1 Neutralization by CD4-Mimetic Miniproteins Revealed by 1.5 Å Cocrystal Structure of Gp120 and M48U1. *Structure*, 21:1018–1029, 2013.

[95] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, 118:11225–11236, 1996.

[96] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics*, 79:926–935, 1983.

[97] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. LINCS: a Linear Constraint Solver for Molecular Simulations. *Journal of Computational Chemistry*, 18:1463–1472, 1997.

[98] W. Kabsch and C. Sander. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, 22:2577–2637, 1983.

[99] S.-H. Xiang, A. Finzi, B. Pacheco, K. Alexander, W. Yuan, C. Rizzuto, C.-C. Huang, P. D. Kwong, and J. Sodroski. A V3 Loop-Dependent Gp120 Element Disrupted by CD4 Binding Stabilizes the Human Immunodeficiency Virus Envelope Glycoprotein Trimer. *Journal of Virology*, 84:3147–3161, 2010.

[100] N. Sullivan, M. Thali, C. Furman, D. D. Ho, and J. Sodroski. Effect of Amino Acid Changes in the V1/V2 Region of the Human Immunodeficiency Virus Type 1 Gp120 Glycoprotein on Subunit Association, Syncytium Formation,

and Recognition by a Neutralizing Antibody. *Journal of Virology*, 67:3674–3679, 1993.

[101] F. Cocchi, A. L. DeVico, A. Garzino-Demo, A. Cara, R. C. Gallo, and P. Lusso. The V3 Domain of the HIV–1 Gp120 Envelope Glycoprotein Is Critical for Chemokine-Mediated Blockade of Infection. *Nature Medicine*, 2:1244–1247, 1996.

[102] P. Lusso, P. L. Earl, F. Sironi, F. Santoro, C. Ripamonti, G. Scarlatti, R. Longhi, E. A. Berger, and S. E. Burastero. Cryptic Nature of a Conserved, CD4-Inducible V3 Loop Neutralization Epitope in the Native Envelope Glycoprotein Oligomer of CCR5-Restricted, but not CXCR4-Using, Primary Human Immunodeficiency Virus Type 1 Strains. *Journal of Virology*, 79:6957–6968, 2005.

A. APPENDIX. SUPERPOSITION OF GP120 CRYSTAL STRUCTURES

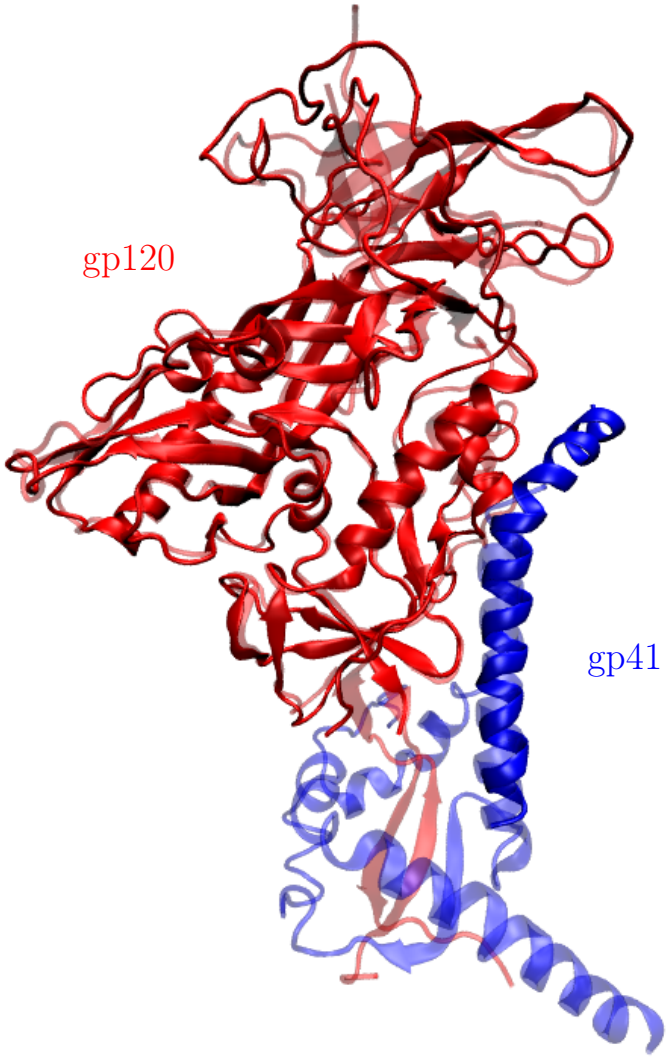


Figure A.1 *Superposition of gp120 and gp41 crystal structures. The bright structure (PDB identifiers: 4NCO [33], 1G9N [93] and 4JZZ [94]) was used in this Thesis. The transparent structure (PDB identifier: 4TVP [57]) that has more resolved residues, especially in gp41, was revealed too late in regard to this Thesis. However, the structure used in this Thesis is relevant because gp120 with its major variable loops was of main interest and those residues did belong to the structure used here. Additionally, the part of gp41 that was needed to link three gp120s to the complex was included in the structure. Nonetheless, the new structure is to be used for an even more accurate model in future work.*

B. APPENDIX. DEVIATION, FLEXIBILITY AND SIZE OF GP120

Table B.1 *The average RMSDs (in unit of nm). The first 200 ns of each simulation was excluded from the calculation.*

Domains	Truncated	Complete	1st	2nd	3rd
core	0.29±0.03	0.29±0.02	0.30±0.01	0.25±0.01	0.36±0.04
core+V4–V5	0.30±0.03	0.31±0.01	0.32±0.02	0.25±0.01	0.36±0.04
core+V3–V5	0.53±0.05	0.41±0.05	0.37±0.02	0.28±0.01	0.52±0.04
core+V1–V5	-	0.57±0.04	0.55±0.02	0.42±0.02	0.63±0.04

Table B.2 *The most flexible residues of gp120 variable loops defined by the average RMSF greater or equal to 0.4 nm. The first 200 ns of each simulation was excluded from the calculation.*

Loop	Truncated	Complete	1st	2nd	3rd
V1	-	143–147, 150		143-150	
V2	-	162–167			
V3	309–315	301–311			
V4				396	
V5	450–451				

Table B.3 *The radius of gyration (in unit of nm) of gp120. The first 200 ns of each simulation was excluded from the calculation.*

Domains	Truncated	Complete	1st	2nd	3rd
Core	2.15±0.02	2.10±0.01	2.09±0.01	2.09±0.01	2.11±0.01
Core+V3–V5	2.20±0.02	2.28±0.03	2.25±0.01	2.26±0.01	2.36±0.01
Core+V1–V5	-	2.45±0.03	2.40±0.01	2.45±0.01	2.59±0.02

C. APPENDIX. MEAN SMALLEST DISTANCES BETWEEN GP120 RESIDUES

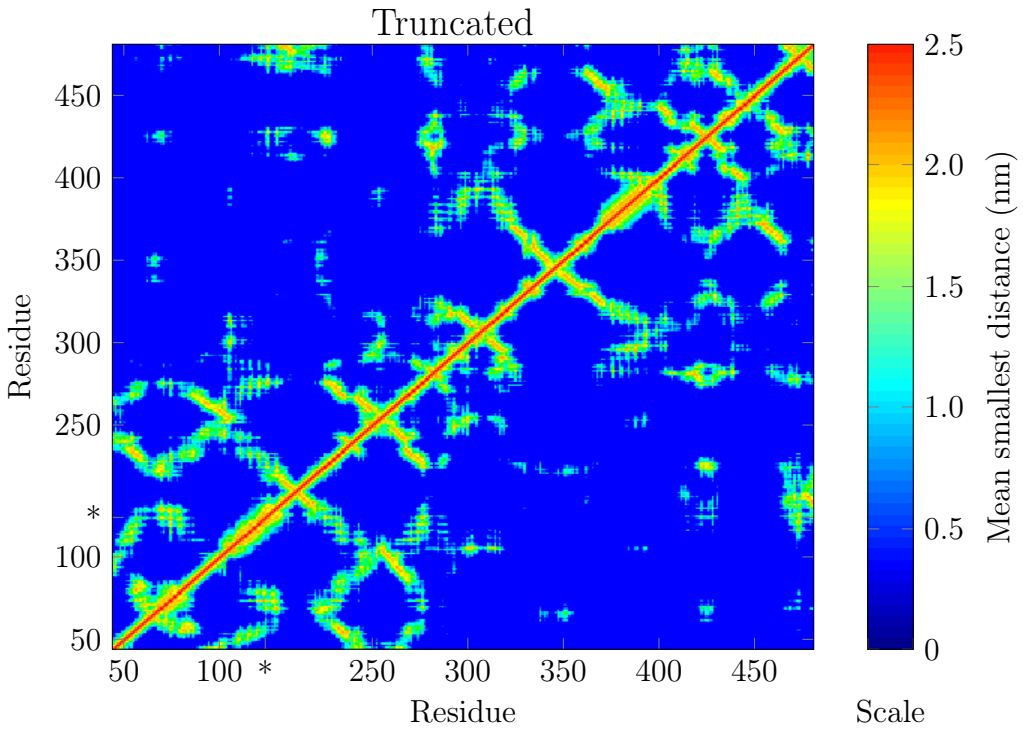


Figure C.1 *The truncated monomer. The mean smallest distance between the residues. The missing residues 124 to 193 are marked with a star (*) in the coordinate system.*

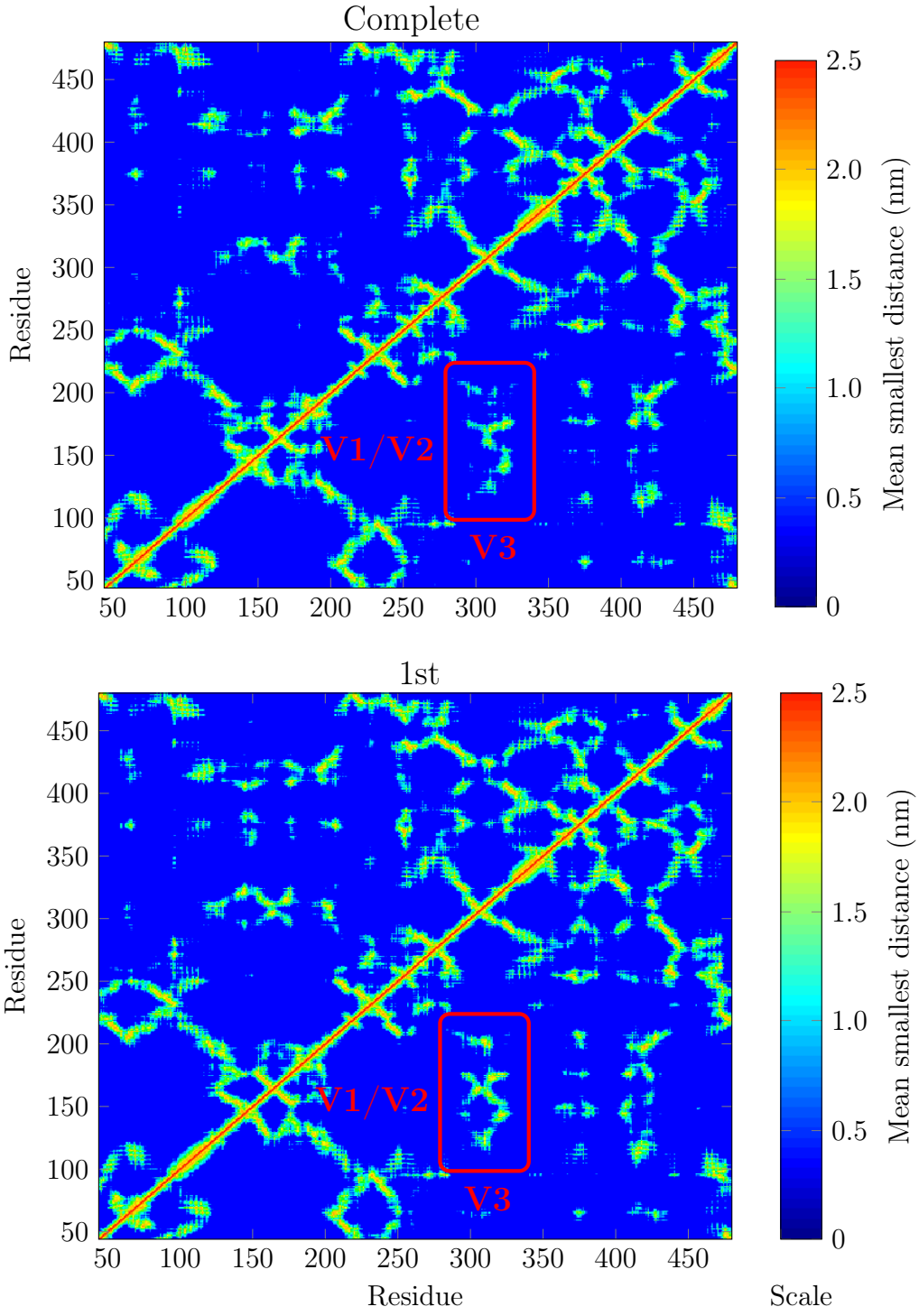


Figure C.2 *The complete monomer and the 1st trimer unit. Mean smallest distance between the residues.*

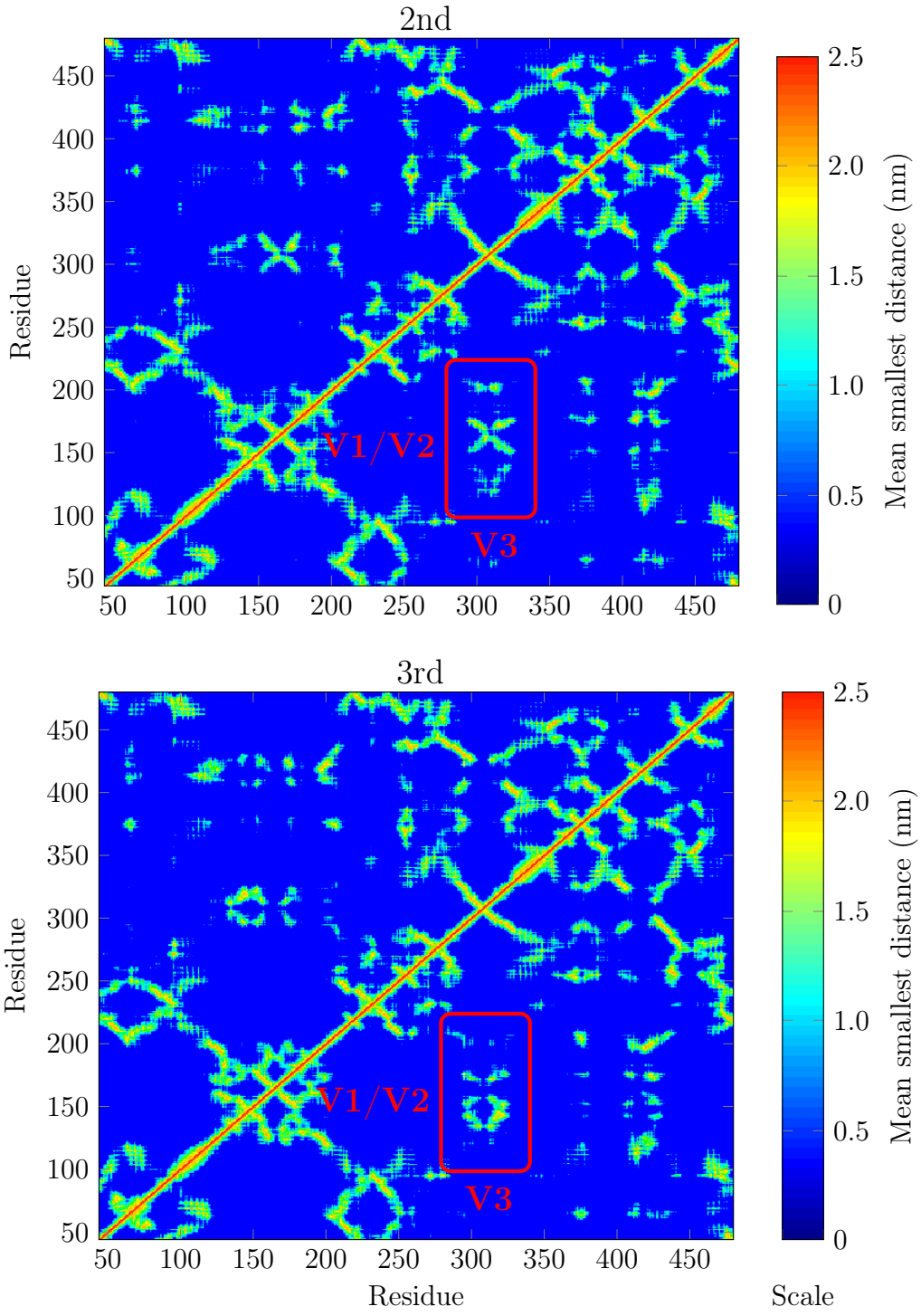


Figure C.3 *The 2nd and 3rd trimer units. Mean smallest distance between the residues.*

D. APPENDIX. HYDROGEN BONDING IN LOOP DOMAINS

Table D.1 *Number of hydrogen bonds between the domains.*

Domains	Truncated	Complete	1st	2nd	3rd
Core–V1/V2	-	6±2	5±2	6±1	7±2
Core–V3	7±2	8±1	8±1	9±1	7±1
V1/V2–V3	-	4±2	4±1	4±1	5±2

Table D.2 *Monomers. Occurrence of the hydrogen bonds (and salt bridges) in time between the V3 domain and the core. The first 200 ns were excluded. All hydrogen bonds existing at least 10 % of the remaining time are shown. Bond marked in bold existed in all systems.*

System	Donor	Acceptor	Occurrence (%)
Truncated	ARG295-Main	ILE431-Main	**38.85**
	ARG295-Side	GLY429-Main	91.21
	ARG295-Side	GLU377-Side	**94.58 (21.80)**
	ARG310-Side	ASP59-Side	11.87 (12.69)
	CYS293-Main	CYS433-Main	**78.35**
	CYS433-Main	CYS293-Main	**59.95**
	ILE408-Main	ARG323-Main	78.22
	ILE431-Main	ARG295-Main	51.84
	THR294-Side	ILE431-Main	**78.89**
Complete	ARG295-Main	ILE431-Main	**43.66**
	ARG295-Side	GLU377-Side	**99.80 (99.79)**
	ASN298-Side	PRO426-Main	19.30
	ASN298-Side	ARG428-Main	19.99
	CYS293-Main	CYS433-Main	**86.47**
	CYS433-Main	CYS293-Main	**87.35**
	GLN430-Main	ARG295-Main	15.97
	GLY429-Main	ASN298-Main	24.32
	ILE408-Main	GLN324-Side	77.10
	THR294-Side	ILE431-Main	**46.22**

Table D.3 *Trimer units. Occurrence of the hydrogen bonds (and salt bridges) in time between the V3 domain and the core. The first 200 ns were excluded. All hydrogen bonds existing at least 10 % of the remaining time are shown. Bonds marked in bold existed in all systems.*

System	Donor	Acceptor	Occurrence (%)
1st	ARG295-Main	ILE431-Main	**27.42**
	ARG295-Side	GLU377-Side	**99.82 (97.53)**
	ARG301-Side	PRO426-Main	36.11
	ARG310-Side	PRO124-Main	23.45
	ARG310-Side	THR199-Side	20.07
	CYS293-Main	CYS433-Main	**85.83**
	CYS433-Main	CYS293-Main	**88.25**
	GLN200-Main	TYR313-Side	76.05
	THR294-Side	ILE431-Main	**34.70**
	TYR313-Side	GLN200-Main	11.06
	TYR313-Side	MET422-Main	78.97
2nd	ARG295-Main	ILE431-Main	**72.89**
	ARG295-Side	GLU377-Side	**99.87 (99.48)**
	ARG301-Side	CYS202-Main	60.04
	ARG301-Side	PRO203-Main	82.81
	ASN297-Side	ARG428-Main	60.76
	CYS293-Main	CYS433-Main	**87.87**
	CYS433-Main	CYS293-Main	**87.77**
	GLN430-Main	ASN297-Main	20.75
	GLN430-Main	ASN297-Side	24.98
	ILE431-Main	ASN297-Side	34.46
	THR294-Side	ILE431-Main	**17.52**
3rd	ARG295-Main	ILE431-Main	**71.70**
	ARG295-Side	GLY429-Main	13.37
	ARG295-Side	GLU377-Side	**96.96 (33.59)**
	CYS293-Main	CYS433-Main	**74.05**
	CYS433-Main	CYS293-Main	**90.46**
	GLN324-Side	GLN410-Side	17.69
	GLN430-Side	PRO296-Main	23.82
	ILE408-Main	GLN324-Side	76.95
	THR294-Side	ILE431-Main	**73.25**
	THR315-Side	PRO426-Main	16.50

Table D.4 *The complete monomer and the trimer units. Occurrence of the hydrogen bonds (and salt bridges) in time between the V1/V2 domain and the core. The first 200 ns were excluded. All hydrogen bonds existing at least 10 % of the remaining time are shown. Bonds marked in bold existed in all systems.*

System	Donor	Acceptor	Occurrence (%)
Complete	ARG407-Side	LEU178-Main	19.14
	ARG407-Side	ASP179-Side	17.80 (5.17)
	ASN159-Main	ASN130-Main	**52.73**
	ASN185-Side	GLU417-Side	10.83
	ASN194-Main	ILE183-Main	43.59
	ASN194-Side	ASP184-Side	34.72
	GLN410-Main	ASN177-Side	59.25
	ILE411-Main	ASN177-Side	30.97
	LEU129-Main	ILE191-Main	64.35
	THR128-Side	CYS193-Main	48.07
	THR195-Side	ASN185-Side	13.20
	TYR188-Side	GLU417-Side	17.26
1st	ARG407-Side	GLU152-Side	27.96 (27.14)
	ARG407-Side	GLU149-Side	60.56 (9.65)
	ASN130-Side	CYS131-Main	34.45
	ASN130-Side	TYR188-Side	10.40
	ASN159-Main	ASN130-Main	**50.39**
	ASN159-Side	THR128-Main	13.46
	ASN159-Side	ASN130-Main	47.67
	ASN177-Side	GLN410-Side	10.79
	ASN194-Main	SER192-Side	10.15
	GLN410-Side	LEU174-Main	24.20
	ILE191-Main	ASN130-Side	13.74
	LEU129-Main	SER192-Main	13.50
	LYS420-Side	ASP179-Side	13.05 (15.77)
	TYR176-Main	GLN410-Side	79.67
2nd	ARG407-Side	GLU152-Side	99.89 (99.84)
	ASN130-Side	TYR188-Main	10.59
	ASN159-Main	ASN130-Main	**89.17**
	ASN159-Side	VAL127-Main	60.36
	LYS409-Side	ASP179-Side	38.07 (50.31)
	LYS420-Side	ASP179-Side	54.96 (68.11)
	TYR176-Main	GLN410-Side	89.25
3rd	ARG407-Side	GLU152-Side	55.19 (54.86)
	ASN130-Side	SER187-Main	33.45
	ASN130-Side	ASP184-Side	44.22
	ASN130-Side	ASP184-Main	44.37
	ASN155-Side	PRO426-Main	16.23
	ASN159-Main	ASN130-Main	**62.03**
	ASN159-Side	THR128-Side	54.18
	ASP184-Main	LEU129-Main	19.28
	ILE183-Main	LEU129-Main	49.83
	THR128-Main	ASN159-Main	42.88
	THR128-Side	ASN159-Main	52.69
	TYR176-Side	TYR423-Main	50.41

Table D.5 *The complete monomer and the trimer units. Occurrence of the hydrogen bonds (and salt bridges) in time between the V1/V2 and the V3 domain. The first 200 ns were excluded. All hydrogen bonds existing at least 10 % of the remaining time are shown.*

System	Donor	Acceptor	Occurrence (%)
Complete	ARG310-Side	GLU171-Side	90.37 (78.61)
	ARG310-Side	GLN169-Side	12.52
	ASN139-Side	GLU317-Main	12.07
	GLN324-Side	TYR176-Main	30.15
	ILE319-Main	THR141-Main	10.95
	ILE319-Main	SER142-Side	21.34
	LEU174-Main	TYR313-Side	24.25
	SER143-Main	GLY320-Main	11.00
	SER143-Side	ASP321-Side	12.83
	TYR172-Side	GLY320-Main	16.78
1st	ASN305-Main	SER163-Main	58.08
	GLN324-Side	TYR176-Side	69.08
	LEU174-Main	THR315-Main	75.17
	SER143-Side	GLY320-Main	65.67
	TYR176-Side	GLY320-Main	21.71
2nd	ARG295-Side	TYR176-Side	48.93
	LEU174-Main	THR315-Main	74.41
	LYS302-Side	GLU171-Side	90.20 (5.77)
	THR314-Side	TYR172-Main	65.81
	THR315-Side	LEU174-Main	63.22
3rd	ARG295-Side	TYR176-Side	15.53
	ASN297-Side	SER142-Main	74.52
	GLN324-Side	GLU152-Side	60.72
	GLU146-Main	ILE318-Main	41.70
	GLU317-Main	THR141-Main	37.39
	GLY320-Main	SER144-Main	59.75
	THR141-Side	GLU317-Main	55.36
	THR315-Main	TYR172-Side	20.42
	TYR172-Side	THR315-Side	23.64

Table D.6 *Trimer units. Occurrence of hydrogen bonds (and salt bridges) in time between the variable loops and their adjacent gp120 units in the trimer. The first 200 ns were excluded. All hydrogen bonds existing at least 10 % of the remaining time are shown. Bonds marked in bold existed in all systems.*

Units	Donor	Acceptor	Domains	Occurrence (%)
1st–2nd	ARG165-Side	ASP166-Side	V2–V2	**32.00 (34.18)**
	ASN305-Side	ILE191-Main	V3–V2	33.09
	ASN305-Side	LEU190-Main	V3–V2	10.76
	CYS193-Main	ASN305-Main	V2–V3	25.78
	GLY307-Main	CYS193-Main	V3–V2	77.06
1st–3rd	ARG165-Side	ASP166-Side	V2–V2	**84.26 (0.00)**
	ARG165-Main	ASP166-Main	V2–V2	77.87 (0.00)
	ARG301-Side	ASP184-Side	V3–V2	50.46 (58.59)
	ARG310-Side	ALA186-Main	V3–V2	21.95
	ARG310-Main	LEU134-Main	V1–V3	50.53
	ARG310-Side	ASP179-Side	V3–V2	44.09 (25.06)
	GLN169-Side	CYS126-Main	V2-core	15.21
	GLN169-Side	THR128-Side	V2–core	25.28
	SER192-Main	GLU171-Side	V2–V2	50.20
	SER192-Side	GLU171-Side	V2–V2	54.75
	TYR188-Main	ALA311-Main	V3–V3	65.44
	TYR313-Main	TYR188-Main	V3–V2	59.15
	VAL168-Main	ARG165-Main	V2–V2	43.58
2nd–3rd	ARG165-Side	ASP166-Side	V2–V2	**34.84 (00.25)**
	ARG165-Main	LEU190-Main	V2–V2	64.89
	ARG165-Side	THR162-Main	V2–V2	21.41
	ARG165-Side	SER163-Main	V2–V2	29.33
	ARG189-Side	ASP166-Side	V2–V2	38.45 (47.92)
	ARG310-Side	GLU417-Side	V3–core	73.25 (82.10)
	ASP166-Main	LEU190-Main	V2–V2	20.59

E. APPENDIX. SECONDARY STRUCTURE OF LOOPS

Table E.1 *The average number of residues with β-structure types on V3 and V1/V2.*

Domain	Structure	Truncated	Complete	1st	2nd	3rd
V3	β-sheet	12±2	7±4	1±2	11±1	9±2
	β-bridge	0	0±1	1±1	0	0
V1/V2	β-sheet	-	7±4	16±1	9±3	14±1
	β-bridge	-	2±2	4±1	2±1	1±1

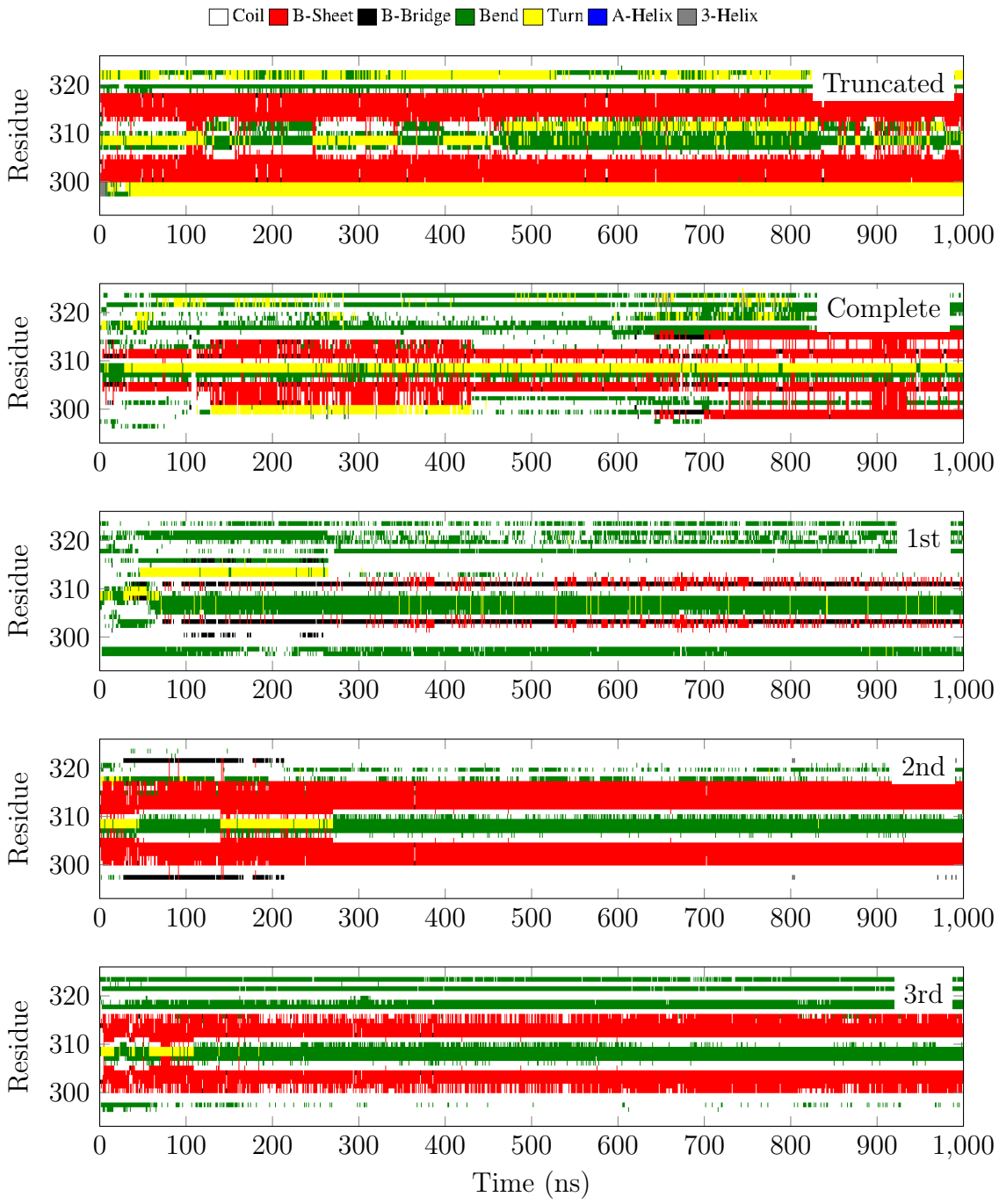


Figure E.1 *Secondary structure of the V3 domain as a function of time in each system.*

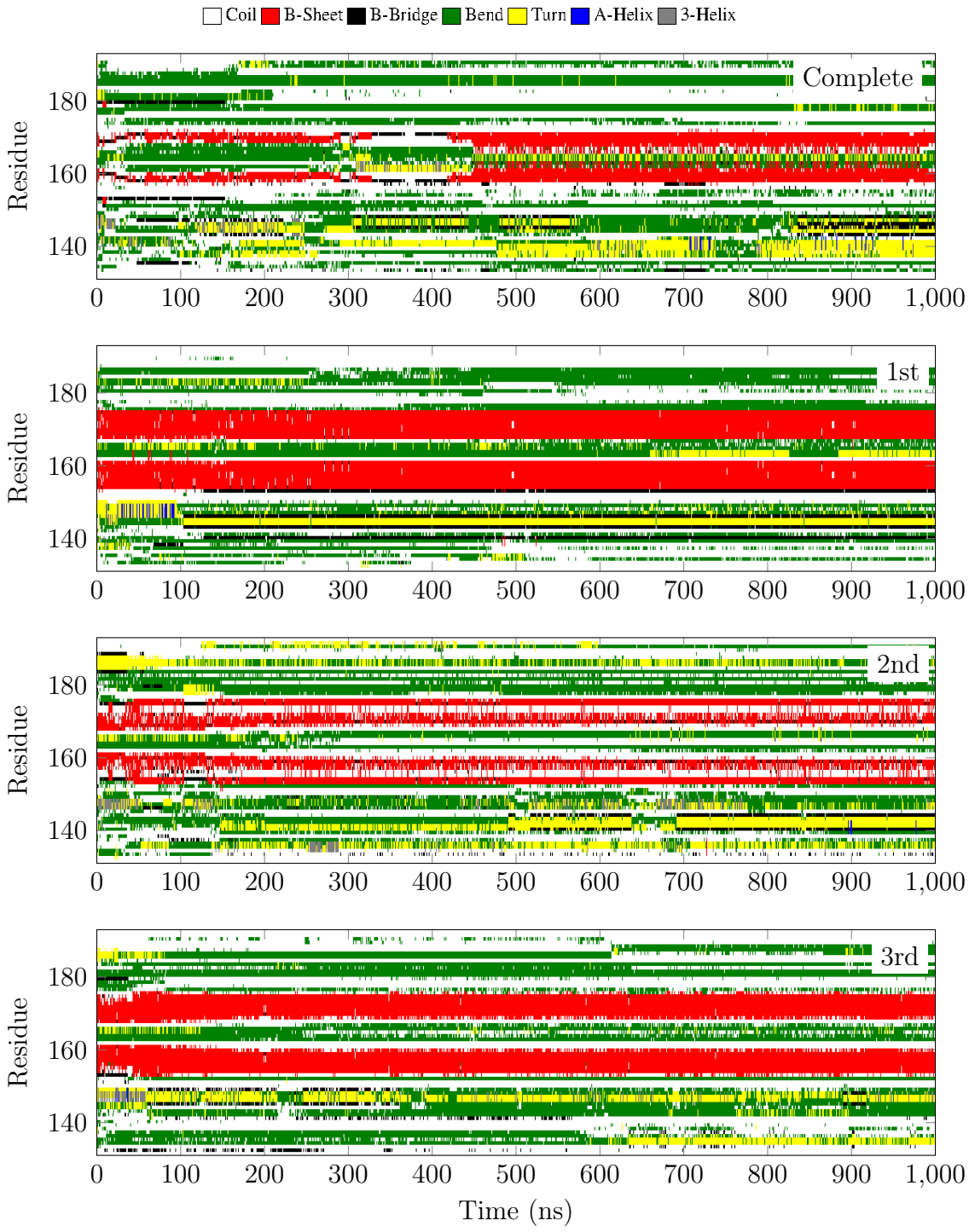


Figure E.2 *Secondary structure of the V1/V2 domain as a function of time in each system.*